Explainable Reinforcement Learning Through a Causal Lens

Prashan Madumal Mathugama Babun Appuhamilage
ORCID: 0000-0002-2466-670X

Submitted in total fulfilment of the requirements of the degree of

Doctor of Philosophy

School of Computing and Information Systems
The University of Melbourne

July 2021

Abstract

This thesis investigates methods for explaining and understanding how and why reinforcement learning agents select actions, from a causal perspective. Understanding the behaviours, decisions and actions exhibited by artificially intelligent agents has been a central theme of interest since the inception of agent research. As systems grow in complexity, the agents' underlying reasoning mechanisms can become opaque and the intelligibility towards humans can be diminished, which can have negative consequences in high-stakes and highly-collaborative domains. The explainable agency of an autonomous agent can aid in transferring the knowledge of this reasoning process to the user to improve intelligibility. If we are to build effective explainable agency, a careful inspection of how humans generate, select and communicate explanations is needed. Explaining the behaviour and actions of sequential decision making reinforcement learning (RL) agents introduces challenges such as handling long-term goals and rewards, in contrast to one-shot explanations in which the attention of explainability literature has largely focused.

Taking inspirations from cognitive science and philosophy literature on the nature of explanation, this thesis presents a novel explainable model —action influence models— that can generate causal explanations for reinforcement learning agents. A human-centred approach is followed to extend action influence models to handle distal explanations of actions, i.e. explanations that present future causal dependencies. To facilitate an end-to-end explainable agency, an action influence discovery algorithm is proposed to learn the structure of the causal relationships from the RL agent's interactions. Further, a dialogue model is also introduced, that can instantiate the interactions of an explanation dialogue. The original work presented in this thesis reveals how a causal and human-centred approach can bring forth a strong explainable agency in RL agents.

Declaration

I, Prashan Madumal, declare that this thesis titled, "Explainable Reinforcement Learning Through a Causal Lens" and the work presented in it are my own. I confirm that

1. the thesis comprises only the candidate's original work towards the degree of Doctor of Philosophy, except where indicated in the preface;
2. due acknowledgement has been made in the text to all other material used; and
3. the thesis is fewer than the 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.

Prashan Madumal

July 2021

Preface

This thesis is submitted in total fulfilment of the requirements for the degree of *Doctor of Philosophy* at the University of Melbourne. the research presented here was primarily conducted at the School of Computing and Information Systems, The University of Melbourne, under the supervision of Prof. Tim Miller, Prof Liz Sonenberg and Prof Frank Vetere.

Below is the list of publications and manuscripts arising from this thesis. I was the principal author of all papers and contributed more than 50% on each paper. I was responsible for designing the algorithms architectures, collecting data-sets, implementation, running experiments and analysing the experimental results. My co-authors provided feedback on the proposed algorithms and models and contributed to the revisions of the manuscripts. Ethics approval to conduct the studies comprising this thesis was provided by The University of Melbourne's human ethics committee (Chapters 3 and 4 - ID: 1953619.1, Chapter 6 - ID: 1647972).

- Part of the contents of Chapter 3 has been published in the following paper: Madumal, P., Miller, T., Sonenberg, L., & Vetere, F. (2020, April). Explainable reinforcement learning through a causal lens. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 03, pp. 2493-2500).
- Part of the contents of Chapter 3 has been published in the following paper: Madumal, P. (2020, April). Explainable Agency in Reinforcement Learning Agents. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 10, pp. 13724-13725).
- Part of the contents of Chapter 6 has been published in the following paper: Madumal, P., Miller, T., Sonenberg, L., & Vetere, F. (2019, May). A Grounded Interaction Protocol for Explainable Artificial Intelligence. In Proceedings of the 18th International Conference on

Autonomous Agents and MultiAgent Systems (pp. 1033-1041).

- Part of the contents of Chapter 6 has been published in the following paper: Madumal, P. (2019, May). Explainable agency in intelligent agents: Doctoral consortium. In Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (pp. 2432-2434).
- Part of the contents of Chapter 4 has been adapted from the pre-print article under review: Madumal, P., Miller, T., Sonenberg, L., & Vetere, F. (2020). Distal explanations for explainable reinforcement learning agents. arXiv preprint arXiv:2001.10284.

I gratefully acknowledge the following sources of funding:

- Google PhD Fellowship
- Melbourne Research Scholarship
- SIGAI/AAAI Travel Grant
- Google PhD Travel Grant
- AAMAS Travel Grant

Acknowledgements

It has been a challenging and rewarding journey that would not have been possible without the support of many people.

First and foremost, I express my gratitude to my supervisors Professor Tim Miller, Professor Liz Sonenberg and Professor Frank Vetere. I am incredibly grateful for your continuous support, guidance and optimism throughout these years. I'm fortunate to have excellent scholars like you as my mentors and this immensely contributed to my success. Tim, thank you for taking me under your wing, for the unwavering support, and for being a great mentor and a friend. You introduced me to the Melbourne coffee culture, and I thoroughly enjoyed the causal encounters we had in all the agent Lab weekly coffee meetups. Tim, you encouraged and gave me the confidence to explore new research directions and expanded my horizons on what's possible. The work presented in this thesis benefited immensely from your guidance. Frank, our discussions were always thought-provoking, with fresh perspectives on the synergies of AI and human factors. I greatly enjoyed your sense of humour, academic curiosity and wisdom. My research journey has undoubtedly been positively influenced by your traits. Liz, your kind-heartedness, as well as far-reaching knowledge in Artificial Intelligence, have been a strong pillar of support in my PhD journey. I much appreciate your vision, insights and encouragement in guiding me.

I have been fortunate to find friendship and companionship among many magnificent PhD scholars. Thank you for all the laughter, random discussions and providing great company. In particular, I like to thank the amazing people of the Agent Lab: Stefan, Steven, David, Daniel, Anubhav, Guang, Lyndon, Micheal, Mor, Michelle, Abeer, Nir and Adrian. It has been humbling to work alongside Ruihan and Fatma, I greatly enjoyed our collaborations and thank you for the support.

I want to thank my brilliant colleagues in the HCI group. Thank you, my neighbours, in the 4.04 office, Nick, Ronny, Unni and Ahed, I greatly enjoyed our frequent conversations. Special thanks to my wonderful friends in the HCI group, Niels, Zhanna, Danula, Senuri, Gabriel, Yousuf, Namrata, Romina, Ebrahim, Fraser, Henrietta, Sarah and many more. I'm glad to have had such great company in my PhD journey. I want to especially thank Joshua and Ronal, for their immense support and collaborations.

Next, I would like to thank the CIS graduate research students members, which I was grateful to be a part of as a committee member. My thanks go to Xin, Difeng, Quishi, Peter and many more. I also want to thank the wonderful PhD scholars of CIS, Tharindu, Sameera, Gayashan, Malinga, Udesh, Amila and many more.

In addition, I want to thank the mentors I met in my PhD journey. David Aha, thank you for being a great mentor in my doctoral consortium, and always being supportive throughout. Michael, thank you for being my mentor through the Google PhD fellowship program, providing insights on industrial research. I am fortunate to have had closely worked with Silvia, Rosina, Mark and more in workshop organising activities and collaborations.

I gratefully acknowledge the financial support I received. This includes the Melbourne Research Scholarship, the Google PhD Fellowship, and travel grants from SIGAI, AAMAS and Google. I further acknowledge all the teachers who guided me. My sincere gratitude goes to the people of Sri Lanka, for supporting me through free education.

I am incredibly thankful for the invaluable support of my family. Thank you, my beloved mother, Nandaseeli and father, Premachandra for giving me the very best and always motivating and encouraging me to define my own path. Thank you to my loving sister Chamindie for always supporting me. Finally, thank you to my Wife, Tharuka. This journey would have not been possible without you. I am blessed to have you by my side.

Thank you!
Prashan Madumal.

Contents

List of Figures

List of Tables

Acronyms

ADF Agent Dialogue Framework.

AI Artificial Intelligence.

BIC Bayesian Information Criterion.

DAG Directed Acyclic Graph.

MDP Markov Decision Process.

RL Reinforcement Learning.

SCM Structural Causal Model.

XAI Explainable Artificial Intelligence.

Chapter 1

Introduction

This thesis concerns itself with explainable reinforcement learning (RL) agents. Inspired by themes prevalent in cognitive sciences, causal explanation models are proposed for reinforcement learning agents. Empirical evaluations in computational and human behavioural studies yield strong results for the causal explanation models.

Understanding the behaviour and the decisions of artificially intelligent machines have been a focal point of Artificial Intelligence (AI) research since its inception [222, 96, 49, 245]. Being intelligible to humans is a highly desirable trait of AI systems, especially in interactive settings. In collaborative environments, to be effective and accurate, agents needs to understand the decisions of others, as is the case for human-human collaborations. The need of intelligible systems is further emphasised in high-stake domains like healthcare, law and finance, where a single action of an agent can have drastic outcomes. The knowledge discrepancy that exists between the user and the AI system in terms of its capabilities, information and reasoning mechanisms can diminish the understanding and the trust of the users. Explainable AI (XAI) aims to bridge this knowledge gap between the autonomous system and its users.

The *explainability* of an autonomous agent can be thought of as the agency it has to communicate the reasoning underlying its behaviour. This can include the capability of providing information that factored into the past behaviour, and information about the decision making *model*, which can help in anticipating future behaviour. The dissemination of such information can consist of different methods such as using transparent behaviour (where the reasoning is embedded and made visible into behaviour itself) and using interpretable models (using intelligible decision-making models). *Explanation* is another facet of the explainable agency of an autonomous system,

where the knowledge transfer is done via explicitly providing explanations of the behaviour. The explanation for some decisions can be in the form of natural language text, visualisations and demonstrations among others.

The high-level reasoning capabilities of autonomous agents have evolved rapidly since the development of Shakey the robot [171], that used predicate logic and symbolic representations of the world to make decisions. In terms of being explainable, these early intelligent systems had the advantage of being inherently interpretable (though perhaps only to their developers) due to having less complex decision-making models. As the capabilities of these agents became more advanced, dedicated explainability facilities also emerged that can support complex decision making. The attention and the importance given to such facilities is evident in the expert systems literature [222, 96, 49, 245, 2, 43]. Autonomous agents' complexity grew at an accelerated pace in the last decade, fueled by the advancements in artificial neural networks. In contrast to early intelligent systems, most modern AI systems have black-box optimisation and decision making systems. While black-box models like neural networks can greatly enhance the performance and the capabilities of an agent, the opaqueness of the decision-making process can reduce the intelligibility of the system.

Reinforcement learning agents are sequential decision making agents that learn behaviour (commonly known as the agent's policy) by interacting with an environment. The policy of the agent is influenced by the rewards obtained through interaction. This makes RL agents fundamentally different from other machine learning techniques like classification, and the explainable agency in RL thus requires a different perspective. The sequential nature of RL also poses a challenge that classification systems do not face. An explainable classifier generally explains a one-shot decision (e.g. why an input was classified into a certain class). In contrast, to be explainable, an RL agent should be able to provide justifications for an action or a sequence of actions, which can depend on the rewards and environmental conditions. Further, though some actions give no immediate reward, they can still be desirable when future rewards are enabled and increased by those actions, making the explainability of RL agents difficult. This thesis focus on the explainability of model-free RL agents, which presents a harder challenge than in model-based RL agents due to the unavailability of a model that can be exploited for intelligibility. Further, generating explanations for agents' actions is a plausible strategy to enable explainable

agency in RL.

The nature of explanation is a well explored notion in philosophy and cognitive science. Having familiar human-centred models of explanation is desirable for explainable systems [131]. Miller's [156] work drew a landscape for explainable AI, taking insights from social sciences, to build successful systems that are accepted by end-users. Indeed, much of the research contributions detailed in this thesis was inspired by philosophy and cognitive science literature in causal and interactive explanations.

With the widespread adoption of autonomous intelligent systems, calls for explainable AI has seen renewed vigour in recent years. This interest in XAI is evident not just in the research community but in popular press [73, 84] and from recent initiatives of explainable AI from tech giants like Google [72] in *responsible AI* guidelines [188] and in AI principles [180]. Contrary to the explosion of AI research, widespread adoption of AI systems has remained cautiously limited due to *ethical* reasons [11] and the lack of *trust* [137] from its users. By building explainable, interpretable and more transparent AI systems, users can potentially be better equipped to understand and therefore learn to trust and better use these systems [156, 99].

1.1 Research Contribution

The problem of Explainable AI can be thought of as a human-agent interaction problem, as it primarily involves transferring the knowledge of *why* an AI model made a decision to a human that is interacting with the system [156]. In most cases, researchers in XAI use their intuition to define what forms a 'good' explanation. Miller, Howe, and Sonenberg [158] argues that this can lead to the failure of the XAI system. The experts that build AI systems are often not in the best position to judge the effectiveness of an explanation to a layperson, which Miller, Howe, and Sonenberg refer to as "Inmates running the asylum". Thus, a human-centred approach to XAI that understand how humans define, generate, select and communicate explanations can arguably form an important research agenda for the XAI community.

Much of the recent progress in explainable AI was concentrated on developing interpretable machine learning methods (e.g. explaining classifiers like convolutional neural networks). Advances have also been in the planning literature [76, 40], in explaining sequential decision making agents in planning and scheduling domains. Comparatively, explainability in reinforcement

learning is underexplored. Much recent work in the explainable RL literature has focused on porting successful interpretable methods like SHAPLEY values and saliency maps from explainable classification and applying them in the RL context [242, 88, 205]. While in theory this makes sense (as most modern RL agents also make use of neural networks) and can help in the intelligibility of the RL agent, in practical scenarios, these methods can be less effective [103]. As [156] emphasised, a grounded approach that takes cues from cognitive science and philosophy may fair better in user-centred settings.

To this end, this thesis aims to look beyond what is *technically possible* to explain and follow a bottom-up approach, first understanding the *nature* of human explanation and then building computational models that encompass that understanding. Seminal work in cognitive and philosophy of science have explored the nature of explanation [53] and how human perceive explanations through a causal lens [248]. In human-computer interaction literature, efforts have been made that follows a user-centric view to understand what constitutes an explanation [241, 1].

The primary focus of the research work will be on developing novel explainable models for reinforcement learning agents, grounded on theories in cognitive science (specifically causal explanation). In addition, another aim of this thesis is to introduce a general explanation dialogue model as an explanation interface. This research will contribute to the growing body of XAI research in a multi-disciplinary manner. Models, methods and findings of this research will potentially be useful not only for the AI community but also for the HCI community. Figure 1.1 visualises the research contribution and the influences of this thesis and how the work is situated in the multi-disciplinary XAI landscape.

1.1.1 Research Questions

This thesis will specifically answer the following research questions (Figure 1.1 shows how individual research questions are positioned in terms of the contribution and the impact).

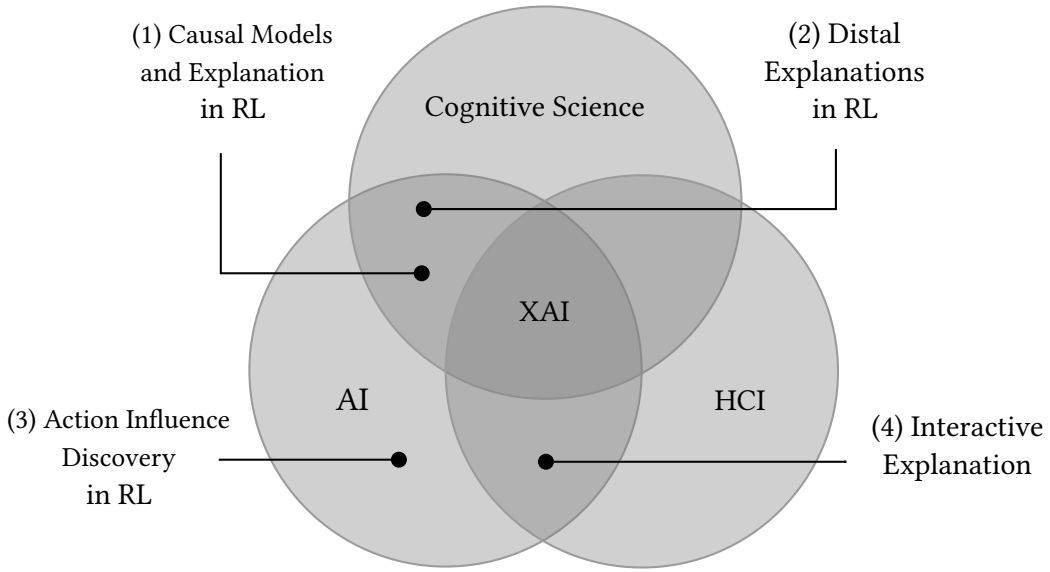


Figure 1.1: Research Agenda of the Explainable Reinforcement Learning.

RQ1: How can reinforcement learning agents provide *causal* explanations of agent behaviour that increase the understanding and trust of the users? (Figure 1.1 (1)).

RQ2: Do *distal* causal explanations of RL agents improve the intelligibility of the agent behaviour? (Figure 1.1 (2)).

RQ3: How can reinforcement learning agents *discover the causal action influence structure* of the domain? (Figure 1.1 (3)).

RQ4: What are the patterns of dialogical explanation that can facilitate *interactive explanations* in XAI systems? (Figure 1.1 (4)).

The central research contribution of the thesis lies in the development of computational and theoretical causal explanation models for reinforcement learning agents.

1.2 Explaining Reinforcement Learning Agents

To address the research questions **RQ1** and **RQ2**, a *causal* explanation approach is followed. This is motivated by the cognitive science theories that point to the usefulness of causal explanations for intelligibility [248, 156, 155]. Using causal explanations as the scaffolding, RQ1 will be

answered by building a computational causal explanation model that can answer 'why' and 'why not' questions about the RL agents actions. This model captures the influences of the agents' actions through the proposed *action influence model* that is based on structural causal models (SCM) [92]. The action influence model contains the state variables, reward variables and the actions of the RL agent. The model can be graphically depicted through a directed acyclic graph (DAG), where the nodes give the state and reward variables with edges annotated with the agent's actions according to the causal influence they have on a particular state/reward variable. Explanations are generated by extracting the *causal chain* of an action that leads to a sink node of the DAG. Further, using causal models enable action influence models to simulate counterfactuals to generate contrastive explanations for why not questions. Action influence models' are evaluated computationally in OpenAI RL benchmarks [31] and in the StarCraft II environment ([235] and Section 1.2.1). To measure the usefulness and the effectiveness of the generated explanations, a human study is conducted using StarCraft II agents, with results showing that the action influence models' explanations perform significantly better than existing local RL explanation methods in metrics; task prediction, perceived explanation quality and trust [110].

RQ2 builds on the insights gained from the evaluation of action influence models in StarCraft II human-agent interaction setting. Participants were asked to provide their own explanations of StarCraft II RL agent behaviour using free text. Through thematically analysing the human explanations of the RL agents' actions, several key concepts were observed. A *distal* explanation model is proposed to implementing these insights primarily focusing on *opportunity causal chains* [156] that explain a causally dependant future action. Importantly, the proposed models and artefacts for RQ1 and RQ2 are model-agnostic, in that, action influence models and distal explanation models do not depend on the different architectures and techniques in the RL paradigm. That is, model-based, model-free, on-policy, off-policy RL algorithms, etc. can all make use of action influence models and distal explanation models as surrogate explainable models to generate explanations.

RQ3 addresses the main shortcoming that exists in action influence and distal explanation models. Both of the aforementioned models require a causal structure that describes how actions influence the variables, to be given in advance (either handcrafted or approximated). The scalability of

handcrafted action influence graphs can suffer in larger domains with complex causal structures. RQ3 answers this shortcoming via learning the action influence structure end-to-end using only the state-action trace data of an RL agent. As before, this method is also model-agnostic and can function across various RL algorithms. *RQ1, RQ2 and RQ3* together form an end-to-end pipeline of an explainable model for RL agents.

RQ4 looks at the explanation interface of an explainable system to propose an interaction protocol in the form of an explanation dialogue model. This dialogue model contains the structure required to maintain a sequential explanation dialogue between the explainer and the explainee. Human explanation dialogues were analysed to create a state model that is formalised using the agent dialogue framework (ADF) [150]. This model can be instantiated to present the causal explanations generated by the action influence and distal explanation models.

Several drawbacks exist in the explainable RL models and the dialogue model proposed in this thesis. For explainable models, it is of importance to infer the knowledge level and the epistemic state of the explainee to provide effective explanations. If the explanations are not given at the correct abstraction level, the intelligibility of the system can fall. The explanation dialogue model can be used as the interaction method to infer the correct abstract level to adjust and select explanations. Further improvements can be made to the explanation dialogue model, to handle different forms of explanation interfaces (e.g. visualisation based) and to have better synergy with sequential decision making agents (where the dialogue flow can be used to identify counterfactuals to generate contrastive explanations).

1.2.1 StarCraft II Environment

Much of the work described in this thesis uses the StarCraft II environment both as a computational benchmark and as a baseline for human-subject evaluations. StarCraft II was coined as a new challenge for artificial intelligence [235], that has a large action and state-space needing macro and micro strategies to compete. In its basic form, StarCraft II is a real-time strategy game, where the player compete against an opponent in an adversarial setting. The objective of the game is to gather resources to build your base of operations while simultaneously destroying the opponent's base. There are units that can be trained using the gathered resources to attack the opponent's base and units. Importantly, StarCraft II agents can exhibit complex strategies

and behaviours, which makes it an ideal test-bed to evaluate explainable models.

Indeed, many research in explainable agents have used StarCraft II environment as sand-box to evaluate and implement explainable agents [181, 179, 143, 89]. The StarCraft II learning environment can provide both pixel-based and state-based information to an RL agent, having the option to use high-level feature maps [235]. The StarCraft II environment also includes a map-maker application. Importantly this allows one to develop their own 'game' with different objectives, goals and rewards. As using the full state and action space is computationally intensive to train an RL agent in a reasonable time frame, the map-maker can be used as a tool that limits the action and state space, and develop interesting scenarios that are better suited for an explainability task. The work described in Chapters 3, 4 and 5 make use of the map-maker application to define and develop custom scenarios. These scenarios include a search and rescue task, a collaborative task (StarCraft II platform can also be used as a human-agent collaboration environment) and a simplified adversarial scenario. These scenarios are detailed at length in Chapter 4.

1.3 Thesis Outline

The remainder of this thesis is arranged as follows. Chapter 2 provides a commentary of related literature that encompass early work in explainable AI, theories of explanation in cognitive and social sciences, explainability in RL agents, causal influence learning and dialogical explanations. The chapter is concluded by drawing insights from cognitive science and social science literature to develop computational models for explainable RL agents.

Chapter 3 presents original work and introduces *action influence model* that answers the **RQ1**, giving a formalisation and definitions of how an explanation looks for a RL agent. Chapter 3 also details on how to generate explanations using action influence models, and report on the computational and human subject evaluation results. Chapter 4 introduces the *distal explanation model* which address the **RQ2**, that improves upon action influence models by generating distal explanations. A similar evaluation methodology is followed in Chapter 4 with new StarCraft II scenarios. Distal explanations are shown to be especially effective in collaborative scenarios in human-subject experiments. Chapter 5 presents original work in action influence discovery, answering the **RQ3**, that addresses the main limitation that exists in action influence and distal

explanation models. To assess the soundness of the proposed action influence discovery method, the generated influence graphs are evaluated against ground truth graphs obtain from the RL simulators. Chapter 6 address the **RQ4**, which discuss the novel interaction protocol as situated in the explanation interface in the broader explainable AI landscape. This explanation dialogue model is developed following a grounded approach that analysed human explanations (instantiated as a state model), formalised using agent dialogue games, and is then evaluated in a human-agent interaction setting.

Following the presentation of the original work, Chapter 7 positions the contribution in the larger explainable AI context and discusses the current and potential future impacts the work can have in the explainable RL landscape. Chapter 7 further summarises the original work and concludes the thesis by presenting a way forward that extends causal explanation models of RL agents to handle abstractions and sequential interactive explanations.

Chapter 2

Related Work

Explainability and explanation are broad concepts that span across multiple disciplines with a vast body of literature going back to the Aristotelian era. To draw a comprehensive path through the literature in the context of explainable agents, this section is structured in a way that gives the reader insight into how foundational concepts in explainability influenced and help navigate the research agenda.

This chapter is arranged as follows; background central to this thesis is first discussed followed by a brief introduction, highlighting the foundations of explanation in philosophy and early work in the AI community. Relevant work in explainable reinforcement learning is then surveyed, followed by a discussion of causal explanation in XAI literature. Then the work of learning causal models are discussed. The dialogical explanation is then surveyed and the chapter is concluded with a summary.

2.1 Background

This section provides a succinct discussion of the paradigms, concepts and evaluation domains that is central to the thesis. Later sections of this Chapter provides commentary on specific research work that encompasses these explainability paradigms and concepts.

2.1.1 Explainability

The development of explainable systems can be categorised into two distinct concepts [89], the *explainable model* and the *explanation interface*. The explainable model deals with transform-

ing the agents' black-box decision-making process into a model that can be used to generate explanations. This can include developing new inherently interpretable models or introducing *surrogate* models that act as proxies having the ability to generate explanations for the underlying decision-making process. Both inherently interpretable models [196] and surrogate models [191, 144, 161] have found success in developing explainable models. Explanation interface developments tend to be user-centric and function as the middle-ware that communicate the generated explanations to the user (explainee). Importantly the explanation interface handles the two-way interaction between the system and the user. This can be in the form of an explanation dialogue. The mode can also differ from text-based, touch-based to visualisation based interactions. These two concepts are akin to the *cognitive process* and the *social process* that humans follow when explaining phenomena [156]. The cognitive process corresponds to the explainable model where facts that form an explanation is generated and the social process is similar to the explanation interface where the goal is to communicate the generated explanations.

The knowledge that is concatenated in an explanation can be structured in several ways. The explanation can be *partial* or *complete*. Complete explanations generally reveal all the causes and facts that led to a particular decision or behaviour of the agent. This type of explanation can contain numerous irrelevant information that had no influence on the end decision of the system [156], and has the risk of inducing a higher cognitive load on the explainee. A compromise can be made between completeness and minimalism by defining a *minimally complete* explanation [122] template for the context, that contain all the necessary causes that influenced the agent's decision. Another dimension for the structure of the explanation lies in *local* vs *global* explanations. As the name suggests, local explanations provide causes for a single decision instance. In classification tasks, this can be in the form of highlighting what features impacted the classification of a single input. In a sequential decision-making agent, local explanations provide justifications for a single action or sequence of actions. Global explanations differ from local explanations in the scope that encompass the explanation model. For instance, in an RL agent, a global explanation would be for the policy instead of for a single action.

Evaluation Metrics

As the definition of explainability is loosely defined and varies across different contexts, the evaluation of explainable AI methods also differs from model to domain. Evaluation methods for XAI can be broadly divided into two phases, computational and human-based evaluations. The former focus on assessing the fidelity and the performance of the developed explainable model. In surrogate explainable models, computational evaluation checks the faithfulness of the model to the original black-box model. Related to this thesis, computational evaluation of explainable RL agents can verify the model through pitting it against the underlying policy and computing the difference.

Human evaluation of XAI systems is a crucial step in assessing the usefulness and effectiveness of the explainable agency. Human evaluations generally take the form of a user study and can contain both quantitative and qualitative metrics. Quantitative evaluations examine the knowledge gained through the explanations by inspecting the mental model of the explainee. Different proxies can be used to infer the updated mental model of the explainee, and a frequently used method for this examination is the task prediction [110]. Here, after providing explanations, the explainee is queried to provide a *prediction* for a new input (explainee would be able to provide a better prediction if explanations successfully made the model intelligible). Qualitative evaluation metrics focus on the qualities of the given explanation from a user perspective. These qualities can include the completeness, satisfaction and sufficiency of the explanation [110]. In addition, qualitative metrics can also be used to evaluate aspects like the trust and the usability of the XAI system. Using both quantitative and qualitative evaluation methods can paint a complete picture of how effective a potential XAI system can be when deployed in an interactive setting.

2.1.2 Reinforcement Learning

The roots of the reinforcement learning paradigm date back to the early days of computer science, neuroscience and cybernetics [118]. Recent years have seen tremendous progress in applying RL methods to accomplish optimisation and strategic tasks [182]. The essence of RL lies in the repeated enforcement of agent behaviour through the reward signals obtained through the interaction with the environment. The standard model of the RL involves solving a problem defined by a finite Markov Decision Process (MDP) to yield the maximum expected reward [219].

The goal of the agent is formalised as the reward signal and the objective of the agent can be then thought of as achieving the goal optimally. The RL agent receives observations from the environment in the form of a state and uses actions to interact. An MDP is defined as a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}\gamma)$, where S and A give state and action spaces respectively (here we assume the state and action space is finite and state features are described by a set of variables ϕ); $\mathcal{T} = \{P_{sa}\}$ a set of state transition functions (P_{sa} denotes state transition distribution of taking action a in state s); $\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is a reward function and $\gamma = [0, 1)$ is a discount factor. Solving this MDP results in a policy π that maps states to actions by maximising the expected sum of rewards received.

Reinforcement learning can be separated into model-based and model-free agents. In model-free RL, agents do not have access to the transition functions $\mathcal{T}$ and the reward function $\mathcal{R}$. Though we focus on model-free RL agents in this thesis, models and methods devised are general enough to be applicable to other sequential decision making agents (e.g. model-based RL, planning agents). Different classes of methods exist that can solve or approximately solve an MDP to obtain the policy for the agent. Temporal difference learning methods like q-learning and sarsa are some of the commonly used techniques in small finite state and action spaces [219]. As most domains in practical settings have large state and action space, approximate techniques like REINFORCE, policy gradient (PG), actor-critic have found success [219]. Based on the environment, RL agents can have deterministic or stochastic policies. Another way that RL agents are categorised is based on how the policy update happens, where methods can be divided by having the update done on-policy or off-policy. Another important concept in the RL paradigm is the dilemma of exploration vs exploitation. The agent needs to balance the exploration or the environment (to potentially receive better rewards) and to continue on the current best policy. This conflict of interest can be addressed through using different action selection policies like $\epsilon - greedy$ and multi-armed bandit algorithms [219]. Section 2.3 gives a detailed account of the related work in explainable reinforcement learning later in this Chapter.

2.2 Explanation

Though 'explainable AI' has been coined as a viable solution that can help understand the ever-evolving inner workings of black-box artificial intelligence models that became popular in

the last decade, its roots can be traced back to the expert era. If one considers the theoretical notion of 'explanation', origins of it can be seen even in the works of Aristotle [128]. This section examines key philosophical and cognitive science foundations of explanation and how they might better inform the research agenda of explainable AI. This section further highlights notable early work in explainable AI, especially in expert systems and applications.

2.2.1 Nature of Explanation

An explanation can be thought of as transferring knowledge of some fact or an event from one party to another. We can define this knowledge transfer as the *act of explaining* [134]. There are many facets of explanation that resides within this act of explaining.

Causal Explanations

The nature of explanation and causality have been tightly coupled in philosophical and cognitive science literature. Aristotle argued that the search for causes for an explanation was a search for answers to the question "why?" [74]. Causality can be separated into two major schools of thought. *Dependence* theories on causality argue that if there is a cause between two events, the second event should always be followed by the first event [133]. Dependence theories of causality can also be extended to understand *counterfactuals*.

In contrast, *Transference* theories are defined on the physical causation of the transference of energy between objects [64]. An event and its causation according to transference theories are defined as the quantity of energy transferred. Counterfactuals are also implicitly present in transference theories as energy transfers that are unnatural.

It is evident from the literature in both cognitive science and philosophy that causality is a major pillar in the foundation of explanation. The school of Dependence theories of causation seems to be better suited in forming the basis in explainable AI work, and indeed this thesis is largely influenced by this school of thought. A more substantial discussion on causal explanations from the perspective of explainable AI is done in Section 2.4 of this chapter.

The Structure of Explanations

Miller [156] argues that the act of explanation can be thought of as a twofold process. First, the *Cognitive process* handles the generation of the explanation, where finding the causes, counterfactuals and abductive reasoning is performed. The selected causes, counterfactuals, facts that constitute the explanation are defined as the *explanans*. The event that explanans are generated for is called the *explanandum* [156]. The resulting explanans from this cognitive process can be thought of as the *product* of an explanation [141].

The second process is the *Social process* of the explanation [156] and describes how the knowledge is transferred between the explainer and the explainee. As Chapter 6 of this thesis discusses, this process can be a two-way dialogue between the explainer and the explainee. The main goal of this process is to convey information to the explainee to understand an event. The request for this understanding can be posed as a *question* to the explainer. It is also important to note that this process can be a continuous interaction between the explainer and the explainee and multiple agents can participate in this interaction.

Though we see a substantial body of evolving work in the XAI community that address the cognitive process of an explanation, less progress has been made towards incorporating the social process of the explanation into XAI.

Explanatory Questions

To begin the cognitive process of the explanation, an explanandum is needed. This is often inferred through a question posed by the explainee (though there are explanations that do not need a question). An explanation can be thought of as an answer to a 'why' question, as explanations are viewed as inherently providing causal information [134]. Though there are counter theses for the causal nature of explanation [189], causality can still be used as a tool to categorise the types of explanatory questions. Pearl and Mackenzie [178]'s ladder of causation is a good candidate to distinguish the types of questions based on the level of causation [178]. Miller separated explanatory questions into three groups based on the ladder of causation.

'What' questions need associative reasoning capabilities in the cognitive process to generate explanations. An explainee might need to know what 'changed' in event A when another event B occurred. On the other hand, the explainee might want to know what will change when event

A is modified to event B or if event B still occur given a change to event A. This type of questions can be categorised as 'How' questions and require an intervention mechanism in the cognitive process to generate explanations. 'Why' questions are the most challenging questions for the cognitive process of explanation. Why questions require counterfactual reasoning and need to examine and simulate alternative causes of an event to measure what happens to the proceeding events. E.g. Instead of event A occurring, what would happen to event B when event C occurs. A why question also implicitly contain a 'why not' question as the goal of the explainee here is to understand the counterfactual event [156].

Abstraction in Explanation

As noted above, an explanation can be thought of as a knowledge transfer. The knowledge or the information in an explanation can be instantiated at different levels that may depend on the explainee and the explainer.

This notion of abstraction is captured well in Aristotle's theory of abstraction and the *Four Causes* model [16, 94]. Aristotle's proposed framework abstracts answer's to four levels. First, the *Material* level, where the substance of something is explained (e.g. rain is made out of water droplets). Second, the *Formal*, where the properties of something is explained (e.g. a water droplet is spherical). Third, the *Efficient*, in which the proximal rules and mechanisms might cause something to change (e.g. wind might distort the shape of a rain droplet). Forth, the *Final*, where the end goal of something is explained (e.g. replenishing water sources can be thought of as the goal of the rain). Thus the same why question can have different explanations (multiple explanations as well) based on these levels.

Dennett's [60] seminal work also proposed a similar model that is based on *physical*, *design* and *intention*. In developing XAI techniques, a suitable abstraction model similar to the above should be adapted as it is clear that questions posed by the explainee can be strongly linked to different levels of knowledge.

Above discussed aspects of explanation from a philosophical perspective can be used as the foundation for developing rigorous XAI techniques. Though recent concepts and work in XAI can be mapped to these foundations, it is also important to correctly merge these concepts as a bottom-up approach in developing XAI techniques.

2.2.2 Early Work in XAI

As machines get more and more complex, the need to understand them also deepens. This is the same in Artificial Intelligence, and the first glimpse of explainable systems can be seen in the expert systems paradigm. This section aims to capture the early work that led to the development of XAI that is visible today.

Applications

Perhaps the best examples of the need for an explainable system can be found in critical domains such as healthcare. It is a necessity that these systems are well understood by medical practitioners. Swartout [222] introduced a medical therapy advisor that had the capability to provide explanations of the systems methods used for the diagnoses and how they are applied. Swartout later introduced the XPLAIN system that aims to provide automatically generated explanations based on goals [221, 223]. The XPLAIN system also has the advantage of providing justification as explanations. Hasling, Clancey, and Rennels [96] also explored how diagnostic systems in the medical domain can be made explainable. Hasling, Clancey, and Rennels generated the explanations by identifying the distinction between the domain knowledge and the strategic knowledge of the system.

Frameworks

Clancey [49] introduced a framework for rule-based expert systems that aims to provide explanations that are based on structural, strategic and support knowledge of the system. Wick and Slagle [245] proposed another framework that implements an explanation facility or a role in an expert system. This approach is similar to news reporting where events are explained as pieces of news. The RATIONALE framework was proposed by Abu-Hakima and Oppacher [2] to explain knowledge-based expert systems by reasoning explicitly. The system works by providing the explanation to itself and improving upon them. Importantly Abu-Hakima and Oppacher also note the causal aspect of the explanation process and how it can be used to construct explanations.

Interactive Systems

The importance of the 'User' as the consumer of an explanation was highlighted in several works in explainable expert systems. Ye and Johnson [250] studied the impact an expert system with an explanation facility can have on the end-user. The authors highlighted how the explanation facility can influence users' confidence and attitude towards the expert system. User modelling for explanations was also given attention in the literature, notably Chandrasekaran, Tanner, and Josephson [43] work explored how to generated explanations based on different abstraction levels.

It is evident that since the inception of explanations in the expert system era, researchers have considered concepts like abstraction and the user perspective important in giving explanations. Though these concepts are paid much less attention in recent work. In critical domains like healthcare, expert systems often had an explanation facility, which highlights the importance of explanations in highly sensitive scenarios.

2.3 Explainability in Reinforcement Learning

This section focuses on the body of literature that explores explainability in Reinforcement Learning (RL) agents. The work that has influenced the explanation and interpretability of reinforcement learning, some of which are central to the original work in this thesis is also discussed.

Explanation in RL agents can be categorised into three classes. First, action-based explanations, where the explanation is generated for a specific action (often called local explanations). Here, the explanation is can be focused and based on the underlying model of the agent (e.g. and MDP). Second, policy-based explanations, where the explanation is for the policy of the agent (sometimes called global explanations in RL context). And Third, where the explanation is a visualisation of the agent's belief or attention. This section does not discuss visualisation based explainable methods of RL further. Much of the work for visualisation based explanations are derived from interpretable supervised learning literature. Instead, the discussion of this section focuses on action and policy-based explanations and their derivatives.

2.3.1 Action Based Explanations (Local Explanations)

Literature that sought to generate explanations for MDP based agents falls into the scope of preceding work on explainable RL. Often, these earlier work provided *local* explanations in that the explanation is for a question about an action of the agent.

The concept of 'relevant variables' in a factored state of an MDP was exploited by Elizalde et al. [69] to generate explanations. Explanations were primarily targeted at human trainees of a system and explanations were built-in and were presented when the operator (trainee) selected an incorrect action. An explanation constitutes a relevant variable that is selected by an expert for each action. Elizalde et al. later extended this work to generate explanations automatically based on the utility that a state variable had on the policy selecting the action [67, 68]. Khan, Poupart, and Black [122] was influenced by the relevant variable explanations and proposed minimally sufficient explanations for MDPs. Here, the long term effects of an optimal action are considered when generating the explanation. Three domain-independent templates were used as the basis of explanations. This thesis later uses one of these templates as a benchmark method in the evaluation section. Relevant variable explanations present a straightforward method of generating explanations from an MDP, though their inability to provide contrastive explanations of counterfactuals remains a weakness. Khan, Poupart, and Black [122] attempted to remedy this by generating contrastive explanations through value-function comparisons. The effect MDP based agents' explanations have on 'trust' was examined by Wang, Pynadath, and Hill [243]. Experiments were carried out to measure trust in human-robot teams influenced by Partially Observable MDP based explanations. As the measurement of trust was self-report, it is unclear whether the trust gain was from actually understanding the system.

2.3.2 Policy Explanations (Global Explanations)

Policy explanations make use of the agent's policy to extract explanations. Explanations can be at the local level or the global level. Global level explanations generally provide an explanation for the whole policy. Some studies suggest that humans are more receptive to global explanations of agents in certain situations [236]. We discuss literature on both global and local explanation methods in this section.

Struckmeier, Racca, and Kyrki [215] introduced a model-agnostic explanation generation method

using agent policies. In cases where the underlying model (i.e. the policy function) is a black box, Struckmeier, Racca, and Kyrki sample the policy of the agent to extract relevant state dimensions. Understanding of the agents' policies was measured in a human experiment and the perceived understanding of the human participant was used as a proxy to show the transparency of the agent.

Policy explanations of an agent generally aim to provide a 'global' interpretation of the agent's behaviour. Hayes and Shah [99] sought to improve the transparency by providing policy level explanations for agent-based robot controllers. These behavioural explanations of the agent are considered as 'summaries' of the agent's policy. Discreet, continuous and multi-agent domains were used to evaluate the generated policy descriptions against expert descriptions and were shown to improve the transparency of the robot. Amir and Amir [9] also aims to summarise the agent's behaviour and introduced the HIGHLIGHTS algorithm. Important states are extracted from the agent's execution trace based on the Q-values. Human-subject experiments showed that participants preferred HIGHLIGHTS summary explanations compared to full policy explanations though in some situations participants' assessments did not always correlate with their confidence. Policy summarisation was also explored in the context of inverse reinforcement learning to investigate if these explanations are viable if there is a discrepancy between the agent's model and the human's mental model.

2.3.3 Reinforcement Learning Agent Explanations

Here we discuss work in recent years that specifically use characteristics of reinforcement learning (e.g. rewards) to explain behaviour. These characteristics can be used to create an approximate model of the agent's policy or model and then generate explanations through using the approximate model.

Tabrez, Agrawal, and Hayes [224] proposed a framework (RARE) that repairs the agent's understanding of the domain reward function through explanation. RARE is especially useful in human-agent collaborative scenarios when the human's reward function of the collaborative task is erroneous. Explanations are given to the collaborators to update their own reward function. Human experiments were conducted to demonstrate the effectiveness of the RARE framework in collaborative tasks. Explanation in the context of interactive reinforcement learning (IRL) has

been studied [77]. This approach uses the instructions given in the IRL process to the agent as representations to generate explanations about the future behaviour of the agent. Evaluated through a human study, this method affirms that when explanations are given in a familiar medium to the human (e.g. using instruction representations), they can yield a deeper understanding of the agent. Waa et al. [236] developed a method that can translate an MDP of a RL agent to an interpretable MDP. This translation model can then be used to generate a *contrastive* policy that can be queried using contrastive questions. A pilot study was carried out to evaluate the method, where the reported findings show that participants preferred the interpretable policy level explanations. Though these explanations were contrastive they were not based on an underlying causal model. Reward decomposition was used by Juozapaitis et al. [117] to generate minimally sufficient explanations, where reward differences were used to provide explanations that answer what action does have an 'advantage' over another. Juozapaitis et al. utilise the nature of the reward structure often present in domains to explain action preferences of the agent.

2.3.4 Decision Tree Policy Explanations

Central to this thesis, here the discussion focus on how interpretability and explainability were achieved through representing agents' policies as decision trees or graphs.

From early work that represented the agent policy as a decision tree using the 'G' algorithm [45], past literature has explored how decision trees can be used to represent and abstract policies of MDPs. Roth et al. [194] proposed a Q-improvement algorithm that builds an abstract decision tree policy for factored MDP based RL agents. Although decision tree policies are claimed to be more interpretable to humans than black box policies, the extent to which this is true is unclear as this work lacks human experiments. Abstract policy graphs have also been used as the basis to generate policy level explanations [227]. A feature importance measure was used to abstract multiple states into an abstract state which is then used to build the policy graph. The interpretability of the graph was evaluated computationally which shows a linear growth of the explanation size against an exponential growth of state-space. Although this implicitly demonstrates the interpretability of the approach, human experiments are needed to understand the effectiveness of the method.

Though the above methods address interpretability to an extent, to the best of our knowledge previous literature has not studied how decision nodes from a decision tree can be incorporated *with causal chains* to provide explanations that are human-centred.

2.4 Causal Explanation

Causality has been intimately coupled with explanation literature in various disciplines from philosophy of science [248, 247, 92] to cognitive science [140] and social sciences [104]. Explanations can be non-causal— e.g. describing the nature of an event—, and a well-studied branch of explanation literature [190, 226, 209]. This section focuses on several key areas in causality and causal explanation, expanding upon how attributes in causality can be used to provide explanations.

2.4.1 Assigning Causality to Explanations

An explanation can be thought of as assigning causal responsibility to an event [115]. There are several ways to structure the explanation so that explanans refers to the causal relationships of the event that is being explained (the explanandum).

Causal Chains and Sufficient Explanations

Miller [156] argues that considering *causal attribution* as the causal explanation is incomplete and causal attribution forms only a part of causal explanation. Causal attribution provides all the causes of an event to the explainer, which might overwhelm the user if the event is complex. This is also acknowledged in XAI literature where methods have been created to *select* only the *minimal* and *sufficient* variables to present as the explanation [122]. When selecting causes, it is important to understand how causes can be arranged in different types of *causal chains*.

A causal chain is a path made up of causes that relates to a set of events [107]. Different types of causal chains can produce distinct explanations for the same question from the explainee. Hilton and John [107] discussed five such types of causal chains.

First is the *Temporal* chain, where the order of the events does not affect the outcome of the distal event. Consider the following example with three events A, B and C. A and B are causally

related to C and are causes of C. The order in which event A and B occur is irrelevant in this type of causal chain. Second is the *Coincidental* causal chain type, where the causal relationship of two events holds only in certain situations (e.g. A is not a cause of B in general but becomes a cause in certain environmental states). The third type is the *Unfolding* chains, where the causal relationships between events hold in general. For example, event A causes B, and B causes C in most cases. Fourth type is named *Pre-emptive* causal chains, where the order of the event occurrence is important. Here, one of the proximal events is the cause of the distal event (e.g. A causes C, B would have caused C if A didn't occur).

Central to the work discussed in Chapter 4, *Opportunity chains* is the fifth causal chain type that Hilton and John [107] propose. Here, a distal event is *enabled* by a proximal event. As an example, event A enables B, and B causes C. Event B will not cause C if it is not enabled by A. In generating sufficient causal explanations, one would select the type of the causal chain together with a minimal set of causes that explain the event.

2.4.2 Contrastive Explanation

There are several objectives of an explanation. Primarily, an explanation helps with knowledge transfer between two parties. This knowledge transferring process can also come in different flavours. While gaining new knowledge can be thought of as the most frequent function, updating existing knowledge models are equally important. Miller [156] argues that 'Why' questions seek to update such pre-existing knowledge models of the explainee. The question 'Why event A' implicitly includes an *contrastive* case, where the explainee's real intention is to inquire why event A occurred instead of event B. We term event B as the contrastive case of event A. As the name suggests, when providing an explanation for the question above, it is important to contrast the two events. To contrast, one would then need to identify the situation where event B occurred instead of A. This is known as finding the *counterfactuals* [139, 102]. Miller suggests that people often ask the question 'Why event A' while leaving out the contrasting case 'B' [156]. Lipton argues that the solution for this is to include the contrastive case 'B' in the question [138], 'Why A rather than B?'. Thus to feasibly identify the implicit contrastive case present in 'why' questions, one can use 'why not' questions giving the contrasting case with the question [156]

Counterfactuals

Why questions in the form "Why A?" has the question "Why A rather than B?" question embedded in them. Event A is referred to as the *fact* and event B is called the *foil* [138]. Here the fact is the actual event that occurred and the foil is the event that did not (the counterfactual). While the literature largely agrees that all 'why' questions contain the counterfactual case [106, 139], inferring the counterfactual case can be intractable in practice. Consider the question 'Why A?' which can contain an infinite number of foils events (e.g. B, C, D...), thus it is important to have constraints within the question that can be used to identify the foil easily.

To constrain and identify the foil, different templates of explanatory questions can be used. Van Bouwel and Weber [230] proposed four types of explanatory questions. Of those, the '*P-contrast*' template is closely related to the nature of why questions asked from explainable systems. A *P-contrast* question takes the form 'Why does object *a* have property *A*, rather than property *B*?'. The counterfactual case (*B*) is explicitly given here. This is similar to 'Why' and 'Why not' questions explored in agent explanations [179]. 'Why not' questions take the form 'Why not *B*' and the actual case (fact) is inferred through systems state. In the context of reinforcement learning, the question would take the form 'Why not action *b*', and the actual action (fact) is inferred from the agent's policy.

To instantiate the counterfactual case, the 'world' in which the counterfactual event occurred needs to be simulated, which gives the causes of the counterfactual event. For example, the question 'Why *A*, rather than *B*?', can include the comparison (contrasting) of the causes of *A* with the causes of *B*. This is akin to explaining the *difference* between those two events [138].It is the causes of *B* that needs to be simulated. To simulate the counterfactuals, a model of the world (often a causal model) is needed.

2.4.3 Models of Causal Explanation

There are several formal models of causation in the literature that aims to capture the causal relationships and their effects [177, 87, 133, 195]. Causal models seek to answer questions about the effect variables have over others and the mechanisms that govern them [176]. We use causal models to draw conclusions about causal connections and this process is known as *causal inference*. Causal models can be deterministic or probabilistic, and structural equations

can be used to describe how variables can depend on its causal predecessors [28]. Though there are many popular parametric and non-parametric methods of modelling causality, this discussion focus on the graphical causal models introduced by Pearl [175] and how it is used for explanation.

Pearlian Causal Models

Pearlian causal models also known as *Pearlian DAGs* are a class of causal graphical models popularised by Pearl [175, 176]. A causal model can be represented by a directed acyclic graph (DAG) and the variables can be described using a structural causal model (SCM) [92]. When the underlying system or environment can be represented by a Pearlian DAG, *causal discovery* can be attempted to learn the causal structure [56]. This is a crucial property used in the work described in Chapter 5.

Halpern and Pearl's [91, 92] definition of *causal explanation* uses SCM's to model counterfactuals and extracts explanations. Halpern and Pearl's formalisation explicitly considers an agent that can autonomously generate explanations (e.g. a planning agent [93]). Variables of the model is divided into two sets; *exogenous* variables, which often exist outside the agent's model with values determined by external factors and *endogenous* variables which contain inside the agent's model with values affected by their relationship with other variables. In SCM's, each endogenous variable's value is determined by a function (a structural equation). This assignment of values to the variables of the causal model is known as the *context* [92]. Halpern and Pearl further expand upon SCM's by providing a criteria for a *actual cause* of an event $X = x$ (endogenous variable X set to the value x) as a set of events W given that following three conditions hold.

AC1: Both X and W are true in the actual world;

AC2: If there exists any counterfactual values for the events W, then X would not have occurred;

AC3: W is minimal, having no irrelevant events.

Halpern and Pearl's formalism has similarities with concepts introduced in explainability literature, notably in agent-based explanations. Minimally sufficient explanations [122] for causal explanations can be derived using SCM's as they satisfy the AC3 condition above. In the context of reinforcement learning agents, a *context* in the SCM can be thought of as the *current state* of the agent. Derivation of explainability concepts and transforming SCM's for the RL agent context forms a central contribution of this thesis and is discussed at length in Chapter 3.

2.4.4 Causal Explainability

The recent boom in the search for explainability also prompted exploration of how causality can be used to improve the understanding of an AI system, though most work has focused on the interpretability aspect.

Chattopadhyay et al. [46] introduced a new neural architecture that views each layer of the network as an SCM and proposes methods to calculate the average causal effect. Narendra et al. [166] followed a similar approach that abstract the neural network as an SCM by applying a function as a filter on each layer of the network. Harradon, Druce, and Ruttenberg [95] introduced a human-centred approach of causal interpretability by auto-encoding neuron activation. The major advantage of this architecture lies in its ability to extract human-understandable 'concepts' while building causal relationships between them. Zhao and Hastie [255] argue that to generate interpretable causal relationships of black-box learning models, domain knowledge has to be encoded as a causal graph with having means to visualise the graph. Causal explanation generation for sequence-to-sequence models was explored by Alvarez-Melis and Jaakkola [6]. This framework is model-agnostic and infers causal dependencies between the input and the output tokens to select a set of explanations. Martinez and Marca [149] used causal models to generate explanations for visual models. They used observational and interventional causal models to produce a counterfactual image as the explanation. Causal frameworks have also been used to understand GANs [19]. Bau et al. introduced an explanation framework that can justify *why* images are generated through a visual GAN. Besserve et al. [23] also sought to explain GANs using causal models. They also generated counterfactual images as explanations through manipulating internal variables of the GAN. Learning the underlying causal model of the data that is used by the black-box model was also suggested as a viable way to produce more

interpretable models for machine learning [249].

Recent literature has widely explored how causal models can be used to make machine learning models more interpretable. Comparatively, causal explanations in agents are hardly explored. Of explainability in agents, causal explanations in reinforcement learning agents remain as an unexplored research area.

2.5 Learning Causal Influence Models

The ability to represent and reason about the world causally is a hallmark of human intelligence [178, 248]. This enables us to solve complex problems by understanding and exploiting the underlying causal mechanisms. Cognitive psychology studies suggest that children learn causal mechanisms and how to discover new causal relationships early in their cognitive development [85]. Causal knowledge is important in several key aspects. Causal relationships can be used to predict and infer future events. Importantly, causal structures can be used to *intervene* on the world to create new events. As evident from the wealth of psychology literature, causality is tightly integrated into how we discover, predict, act and explain in the world.

2.5.1 Introduction

Advances in deep networks brought about a new wave of achievements in learning agents [13]. In many cases, these agents learn associative relationships between inputs and outputs and while being proficient in one task, often fails to generalise further. The complexity of the models of these agents learns also hinders interpretability. Incorporating some form of causality into these agents can potentially yield better performance and generalise better to novel situations [21].

Causal mechanisms can be introduced to agents in several different ways. Similar to humans, agents can be trained to learn the causal relationships between variables which we call *causal discovery*. In some instances, variables available to the agent might not constitute an apparent causal structure (e.g. pixel input). This problem can be alleviated through *causal representations*, where the raw input can be abstracted to generate a representation that has causal relationships. Lastly, the agent can use a causal model to better inform the decision-making pipeline using a *causal reasoning* approach.

2.5.2 Causal Discovery

Learning the underlying mechanisms that guide the occurrences of events is the first step in the causal ladder. Causal discovery techniques investigate which variable's change of value would *influence* or *cause* another variable's value to change. In contrast, *causal inference* methods make use of such existing relationship to see to what extent would a variable change when another is modified. We do not discuss causal inference further in this thesis, and mainly focus on the causal discovery literature.

Causal discovery methods are widely studied in different domains and in the artificial intelligence literature as a whole. We focus our attention on work that uses agent-oriented learning methods here. These works can broadly be divided into three areas; score based methods, constraint based methods and neural based methods. Although the gold standard in causal discovery is randomised control trial techniques (a form of intervention), in practice interventional methods are harder to perform due to limitations in repetition (e.g. in medical domains). Note that most learning agents do not have this limitation as they are trained often in a simulated environment.

2.5.3 Score based Methods

Score-based methods assess the validity of a causal graph $\mathcal{G}$ using some score function S. This score can be computed from observational or interventional data and can be used to find the best graph by searching through the space of possible directed graphs in the space.

Schwarz et al.'s Bayesian Information Criterion (BIC) [204] is one of the popular methods that is used for model selection. The model with the lowest BIC is selected using a likelihood function. Another common scoring method is the Minimum Description Length (MDL) [48], where the main concept is to select the model that has the shortest description of the data. Geiger and Heckerman extended scoring functions to handle continuous variables by introducing the Bayesian Gaussian equivalent score [79]. Heckerman, Geiger, and Chickering also introduced a Bayesian approach for scoring called Bayesian Dirichlet equivalence score [100] based on the likelihood equivalence.

Another class of score-based methods rely on greedy search to explore the space of possible causal graphs. A well known greedy score-based method is the Greedy Equivalence Search (GES) [153]. The scoring happens at the node level, where a node is chosen and all the neighbours

are scored, and one with the highest score being chosen as the next node if the overall score of the graph improves. Several variations and extensions have been suggested for GES. Hauser and Bühlmann [97] introduced Interventional Greedy Equivalence Search (IGES), by step-wise modification of the graph to allow enhanced estimation. To improve the performance of GES, Ramsey et al. [186] proposed the Fast Greedy Equivalence Search (FGES), which can handle millions of variables. FGES introduced parallelism and allowed the graph to disregard the Markov factorisation.

2.5.4 Constraint-based Methods

Constraint-based approaches construct a directed graph $\mathcal{G}$ that satisfy a given set of constraints. Often many methods use conditional independence of the data distribution to apply constraints that can narrow down the candidate graph that describes the data.

One of the best known constraint-based methods is the Peter-Clark (PC) algorithm [214], which relies on the *faithfulness* criterion, where all the independent variables in a directed graph need to satisfy the d-separation [80, 98] (where the separation of the set of variables is assessed). PC algorithm is dependent on the order in which the algorithm considers each variable. Colombo and Maathuis [50] introduced PC-stable to mitigate this by introducing a queue that saves nodes before removing them. More recently, Tsagris [228] proposed another improvement to the PC algorithm and introduced MPC by adding a new rule to the original where cycles would be prevented. Another algorithm, parallel-PC was also recently proposed by Le et al. [132], with the main difference being having parallelism when conditional independence test was done. PC-simple was developed to handle high-dimensional data by Bühlmann, Kalisch, and Maathuis [34], where only the strongly related variables would be analysed. Aliferis, Tsamardinos, and Statnikov [5] proposed the HITON-PC algorithm, another variation of the PC algorithm combining features of PC-simple and PC-stable. PC algorithm has also been extended to deal with temporal data, like the Dynamic Online Causal Learning algorithm proposed by Kummerfeld, Danks, and Cognition [127].

Fast Causal Inference (FCI) method is another widely used causal discovery algorithm introduced by Spirtes et al. [214]. Importantly, FCI disregard causal sufficiency [201], where some common causes might not be measurable from data. In contrast to most causal discovery methods that

build a directed acyclic graph (as the causal model), FCI builds a maximal ancestral graph [192]. FCI works first by searching for conditional independence for every pair of variables to create the structure of the graph and then orienting the connections (arrows) of the structure. Several improvements have been suggested for FCI such as Really Fast Causal Inference (RFCI) developed by Colombo et al. [51]. RFCI does not perform conditional independence tests for all the d-separation sets, thus making it significantly faster than the FCI algorithm.

2.5.5 Neural Network based Methods

Recent advancements in neural networks have brought forth a new class of causal discovery algorithms, where the computation of the causal graph is approximated through a neural network. Neural network approaches have been successful in finding causal relationships in observational, temporal data.

Convolutional Neural Networks (CNN) [126] have been successfully utilised to infer causal relationships from temporal data. Nauta, Bucur, and Seifert [167] proposed an attention-based mechanism named Temporal Causal Discovery Framework (TCDF) that learns 'temporal' causal graphs. TCDF has the added advantage of having the ability to discover hidden confounders. A recent approach by Marcinkevičs and Vogt [148] also focuses on the causal discovery in temporal data by employing neural nets that can detect Granger causality.

Graph Neural Network (GNN) [199] based approaches have also been proposed for causal discovery. Yu et al. [252] argued that GNNs can be used to approximate the directed graph by employing a variational autoencoder architecture. Another graph encoder based method was introduced by Ng et al. [169], that can learn structural equation models and can handle both discrete and vector-valued variables. A score based method that uses a neural network to approximate the non-linear relationships between variables was proposed by Lachapelle et al. [130], which is based on the continuous optimisation approach presented by Zheng et al. [256].

Generative and Adversarial based approaches have also been suggested to discover causality. Goudet et al. [86] used a generative neural model (CGNN) to learn the causal relationships of the data. CGNN can identify the causal hypothesis, conditional independence and the directed graph of the observational data. Adversarial training has been used to learn causal generative

models [119], where a game-theoretic approach is used to conditionally estimate the variable distribution.

Neural combinatorial optimisation based approaches have also found success in uncovering causal relationships from data. In particular, central to this thesis, reinforcement learning-based approaches were also proposed as a mechanism to search for the correct causal graph. Zhu, Ng, and Chen [257] uses an encoder-decoder model and BIC scoring to assess the causal graph, while using reinforcement learning to search through the graph space. A similar approach was proposed by Huang et al. [113] for incomplete observational data that uses an encoder integrated with reinforcement learning to find the graph using currently available information.

2.6 Dialogical Explanation

Novel explainable and interpretable models have managed to occupy the vast majority of the literature in the most recent rejuvenation in Explainable AI research. Miller argues that researchers should pay equal attention to how the explanation is communicated to the end-user. This *social* process of the explanation handles the interaction that occurs between the explainer and the explainee. The importance of this interaction was rightfully acknowledged by early explainable research that explored explanatory interaction [38, 37, 163]. This section discusses interactive and dialogical explanations both from explainability literature in AI and social sciences, to draw conclusions on how a model of interactive explanation can form a useful component in explainable systems.

2.6.1 Interactive Explanation

Human explanations can provide an anchor to ground the development of interactive explanatory protocols and models. In Cawsey's [37] EDGE system, human explanatory discourse was analysed to propose a user-centred system that can generate explanations in an interactive setting. Cawsey importantly highlights how the user's epistemic knowledge needs to be tracked to provide useful explanations. Slotnick and Moore's [211] QEX explanation model also uses explanatory dialogue to interface with the user and handle user queries. The dialogue history is kept and is used to inform explanation generation as not to repeat previously provided knowledge. Explanatory dialogue is also useful to provide the context and can help generate

relevant explanations. Mittal and Paris's [159] make use of the explanation discourse of the user to infer the context and discuss implications of having such context in text planning scenarios.

The necessity of a dialogue to the explanation process was emphasised in Moore and Swartout's [163] work. In a natural dialogue, users are free to ask follow-up questions to an explanation, but this can pose difficulties for the explainable system. Moore and Swartout developed a planning system named Explainable Expert System (EES) that took user model and dialogue history into consideration, and uses a pointing system to interact with end-users. Another explanation dialogue planning system was proposed by Suthers, Woolf, and Cornell [218] that is interactive and contextual. Suthers, Woolf, and Cornell further discuss how there are discrepancies between cognitive theories of explanation and AI models of explanation. Their explanation planning architecture is sensitive to the dialogue history of the explainee and the user model, which enables incremental explanations. Explanation history is a useful heuristic as noted above and can also serve as a knowledge base to provide future explanations. Rosenblum and Moore [193] used case-based reasoning on explanatory dialogue history to extract contextual effects, and implemented an intelligent tutoring system that uses the developed computational model.

Another critical aspect of explainable systems is their ability to influence the trust a user can have towards the system. Explanatory dialogues can improve and maintain user trust as it allows users much longer interaction windows. Nothdurft, Heinroth, and Minker [173] studied the effects explanation dialogues can have on user trust in domain-dependent tasks which indicated the significance of interactions in maintaining the trust of the users. Nothdurft et al. [172] also highlighted how adaptive explanatory dialogue can prevent the loss of trust in cooperative tasks.

2.6.2 Dialogue Models

Generating and communicating explanations interactively can be a daunting task for explainable systems. The structure and the flow of this interaction between the explainer (the system) and the explainee (the user) need to be modelled accurately in order to provide such sequential explanations. Dialogue models are a popular way of modelling this interaction. Dialogue models can further provide a formalisation to the act of explanation. Though dialogue models usually refer to verbal or textual modes of communication, the same concepts are applicable in building other modes of interactions (e.g. visual-based, gaze-based and touch-based)

Dialogue models are widely explored in the philosophy, social science and cognitive psychology literature. To accommodate the communication aspects of explanations, several dialogue models have been proposed. Walton [238, 237] introduced a shift model that has two distinct dialogues: an explanation dialogue and an examination dialogue, where the latter is used to evaluate the success of an explanation. Walton draws from the work of Memory Organizing Packages (MOP) [200] and case-based reasoning to build the routines of the explanation dialogue models. This dialogue model has three stages: opening, argumentation, and closing [238]. Walton suggests an examination dialogue with two rules as the closing stage. These rules are governed by the explainee, which corresponds to the understanding of an explanation [239]. This sets the premise for the examination dialogue of an explanation and the shift between explanation and examination to determine the success of an explanation [237].

A formal dialogical system of explanation was proposed by Walton [239], having three types of conditions: dialogue conditions, understanding conditions, and success conditions. Arioua and Croitoru [12] formalised and extend Walton's dialectical system by incorporating Prakken's [183] framework of dialogue formalisation. Arioua and Croitoru's dialogue architecture also had a shift model that gave the explainee options to challenge the given explanation.

Gilbert note that both the structure and the explanation strategy is important when one builds a formal dialogue model. Recent work of Attari, Heckmann, and Schlangen [15] has focused on annotation-based verbal dialogue models, where the dialogue is represented through a tuple with the questions, information, answers and participants, for collaborative tasks. Other dialogue work has focused on formalising the dialogue structure by introducing a categorisation [4, 3] mechanism for different dialogue acts. Explanation dialogues have also been used as an aid to rectify imprecise planning problems and improve their solutions, where the dialogue was formalised using a planning approach [65].

2.6.3 Argumentation in Explanation

The natural dialogue of the explanation process can often contain an argumentation sub-dialogue [24]. Intuitively this allows the explainee to ask follow-up questions about a given explanation or event contest the explanation through an argument. The purpose of the argumentation dialogue is to resolve knowledge discrepancies between the explainer and the explainee,

which can result in a deeper understanding of the system that is being explained.

Walton and Bex [24] introduced a dialogue system for argumentation and explanation that consists of a communication language that defines the speech acts and protocols that allow transitions in the dialogue. This allows the explainee to challenge and interrogate the given explanations to gain further understanding. Villata et al. [234] focused on modelling information sources to be suited in an argumentation framework, and introduce a socio-cognitive model of trust to support judgements about trustworthiness.

Argumentation dialogues have also received attention explicitly in the Explainable AI community. Sklar and Azhar [208] introduced an argumentation framework for the purpose of explaining agents. The framework was formalised as a dialogue game having inquiry, information-seeking and persuasion dialogue game types. A similar argumentation-based approach for explainable systems was proposed by Zeng et al. [253], where assumption-based argumentation is used to formalise the explanations. Efforts have also been made towards using argumentation in explainable systems deployed in clinical settings, where argument schemes are supplemented as explanations that support a wellness consultation scenario between a human and an agent [198].

This previous work on explanation dialogues is largely conceptual and involves idealized models, and mostly lacks empirical validation. In contrast, the work described in the Chapter 6 take a grounded, data-driven approach to determine what an explanation dialogue should look like.

2.7 Conclusion

The need for explainability in intelligent systems was a well understood problem, with the literature spanning back to the expert systems era [43, 163, 38]. Much of the recent work in explainability was driven by the urgency that comes with having complex black-models making decisions, sometimes in high-stake situations [90]. Though there are numerous sub-areas in explainability, this survey specifically focused on the explainability of reinforcement learning agents, causal explanation and on interactive explanation.

2.7.1 Influence from Social Sciences

De Graaf and Malle [57] commented on how humans would expect familiar models and methods of explainable from artificially intelligent agents. Miller's [156] work further emphasised this fact, and through a comprehensive survey of social and cognitive science literature, drew a landscape for social science inspired explainability. In fact, a major recurring theme in cognitive science literature is the notion of causal models and causal explanations [248]. Causality in artificial intelligence is a prominent research area [176], though causality and causal explanations for the purpose of explainability remains largely under explored. This is especially true in the case of explainable reinforcement learning agents, where there's a distinct lack of causal explanation methods in the literature.

2.7.2 Explaining Reinforcement Learning Agents

Reinforcement learning poses a different explainability challenge than explainability in classification or planning as the agent actively interact with the environment influencing change. Several approaches have been proposed to tackle the explainable reinforcement learning problem such as, relevant variable explanations [122], policy explanations [99] and policy summarisation [9]. Further improvements can be made for human acceptance and understandability by incorporating concepts of explainability from cognitive sciences.

The **RQ1** poses the question "*How can reinforcement learning agents provide* causal *explanations of agent behaviour that increase the understanding and trust of the users?*". Introducing causal explanations to RL agents is a plausible direction that can help to build a familiar model of explainability that can address this research question. Having a causal model of explanation can further help with generating contrastive explanations using counterfactuals [155]. To investigate the **RQ2**, "*Do* distal *causal explanations of RL agents improve the intelligibility of the agent behaviour?*", a human-centred approach to causal explanations, as noted in the cognitive science literature discussed above, can be better suited. **RQ3** focus on "*How can reinforcement learning agents* discover the causal action influence structure *of the domain?*", to address the need for an action influence discovery algorithm to facilitate causal explanation models.

2.7.3 Interactive Explanation

The social process of the explanation —communicating the explanation— is an equally important component of an explainable system alongside the cognitive process. Recent explainable developments have given much less attention to this aspect comparatively though the importance of it was rightfully highlighted by many researchers [156, 57, 89]. Walton's [238, 239, 237] seminal work in dialogical explanation can pave a path for interactive explanations in explainable systems, though a grounded, empirically tested model can be better suited when implemented in intelligent agents. Further, for an intelligent agent to provide interactive explanations, a proper conceptualisation and formalism of the dialogue structure that is transferable to an agent readable format is needed. The **RQ4** aims to mitigate this research gap by addressing "*What are the patterns of dialogical explanation that can facilitate* interactive explanations *in XAI systems?*"

Chapter 3

Causal Explanations in Reinforcement Learning Agents[1]

Many prominent theories in cognitive science propose that humans understand and represent the knowledge of the world through causal relationships. In making sense of the world, we build *causal models* in our mind to encode cause-effect relations of events and use these to *explain* why new events happen by referring to counterfactuals — things that did not happen. This chapter discuss how causal models can be used to derive causal explanations of the behaviour of model-free reinforcement learning agents. We present an approach that learns an *action influence model*, which is an extension of structural causal models (SCMs) [92] during reinforcement learning and encodes causal relationships between variables of interest. This model is then used to generate explanations of behaviour based on counterfactual analysis of the causal model. The explanation module is computationally evaluated in 6 domains, measuring performance and task prediction accuracy. The model is also evaluated in a study with 120 participants who observe agents playing a real-time strategy game (Starcraft II) and then receive explanations of the agents' behaviour. The human evaluation investigates: 1) participants' understanding gained by explanations through task prediction; 2) explanation satisfaction and 3) trust. Results show that causal model explanations perform better on these measures compared to two other baseline explanation models.

[1]This chapter is adapted from the published article: "Explainable reinforcement learning through a causal lens." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 03. 2020.

3.1 Introduction

There is a wealth of pertinent literature in cognitive psychology that explore the nature of explanations and how people understand them. As humans, we view the world through a causal lens [210], building mental models with causal relationships to act in the world, to understand new events and also to *explain* events. Importantly, causal models give people the ability to consider *counterfactuals* — events that did not happen, but could have under different situations. Although this notion of causal explanation is also backed by literature in philosophy and social psychology [104], causality and counterfactuals are only just becoming more prevalent in XAI. Further, compared to the burst of XAI research in supervised learning, explainability in model-free reinforcement learning is hardly explored.

We introduce an *action influence* model for model-free reinforcement learning (RL) agents and provide a formalisation of the model using structural causal models [92]. Action influence models approximate the causal model of the environment relative to actions taken by an agent. Our approach differs from previous work in explainable RL in that we use causal models to generate *contrastive* explanations for *why* and *why not* questions, which previous models lack. Given assumptions about the direction of causal relationships between variables, during the policy learning process, we also learn the quantitative influences that actions have on variables. Which enable our model to reason approximately about counterfactual states and actions. We define how to generate explanations for 'why?' and 'why not?' questions from the action influence model. We define *minimally complete* explanations taking inspiration from social psychology literature [151].

We computationally evaluated our approach on 6 RL benchmarks domains using 6 different RL algorithms. Results indicate that these models are robust and accurate enough to perform task prediction [110, p.12] with a negligible performance impact. We conducted a human study using the implemented model for RL agents trained to play the real-time strategy game *Starcraft II*. Experiments were run for 120 participants, in which we evaluated the participants' performance in task prediction, explanation satisfaction, and trust. Results show that our model performs better than the tested baseline, but its impact on trust is not statistically significant.

The main contribution of this chapter is twofold: 1) We introduce and formalise the *action influence* model based on structural causal models and present definitions to generate explanations;

2) We conduct a between-subject human study to evaluate the proposed model with baselines.

3.2 Related Work

In this section, we briefly discuss the literature that is most closely related to the chapter. There exists a substantial body of literature that explores explaining the policies and actions of Markov Decision Processes (MDP), though most of them do not explicitly focus on reinforcement learning. Elizalde et al. [68] generated explanations by selecting and using 'relevant' variables of states of factored MDPs, evaluated by domain experts. Taking the long term effect an action has, Khan, Poupart, and Black [122] proposed generating sufficient and minimal explanations for MDPs using domain independent templates.

Policy explanations in human-agent interaction settings have been used to achieve transparency [99] and provide summaries of the policies [9]. Explanation in reinforcement learning has been explored, using interactive RL to generate explanations from instructions of a human [77] and to provide contrastive explanations [236]. Soft decision trees have been used to generate more interpretable policies [52], and reward decomposition has been utilized to provide minimum sufficient explanations in RL [116]. However, these explanations are not based on an underlying causal model.

Other work on causal explanation has focused on scientific explanations [197] and explanations using causal trees [170]. Although some recent work has emphasized the importance of causal explanation for explainable AI systems [156, 155, 145, 146], work on generating explanations from causal explanation models for MDPs and RL agents have been absent.

3.3 Causal Models for Explanations

In this section, we introduce the *action influence model*, which is based on *structural causal models* of Halpern and Pearl [92]. For the purpose of implementing RL agents for explanation, we use a scaled-down version of the full Starcraft II 1v1 match (an adversarial scenario) with 4 actions and 9 state variables for the agent's model (see Figure 3.1). In the following sections we use this Starcraft II scenario accompanied by Figure 3.1 as our running example.

3.3.1 Preliminaries : Structural Causal Models

Structural causal models (SCMs) [92] represent the world using random variables, divided into exogenous (external) and endogenous (internal), some of which might have causal relationships which each other. These relationships can be described with a set of *structural equations*. Formally, a *signature* S is a tuple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$, where $\mathcal{U}$ is the set of exogenous variables, $\mathcal{V}$ the set of endogenous variables, and $\mathcal{R}$ is a function that denotes the range of values for every variable $\mathcal{Y} \in \mathcal{U} \cup \mathcal{V}$.

Definition 3.3.1. A *structural causal model* is a tuple $M = (\mathcal{S}, \mathcal{F})$, where $\mathcal{F}$ denotes a set of structural equations, one for each $X \in \mathcal{V}$, such that $F_X : (\times_{U \in \mathcal{U}} \mathcal{R}(U)) \times \left(\times_{Y \in \mathcal{V} - \{X\}} \mathcal{R}(Y)\right) \rightarrow \mathcal{R}(X)$ give the value of X based on other variables in $\mathcal{U} \cup \mathcal{V}$. That is, the equation F_X defines the value of X based on some other variables in the model. ■

A *context* $\vec{u}$ is a vector of unique values of each exogenous variable $u \in \mathcal{U}$. A *situation* is defined as a model/context pair $(M, \vec{u})$. An *instantiation* is defined by assigning variables the values corresponding to those defined by their structural equations. An *actual cause* of an event φ is a vector of endogenous variables and their values such that there is some counterfactual context in which the variables in the cause are different and the event φ does not occur. An explanation is those causes that an explainee does not already know. For a more complete review of SCM's we direct the reader to [92].

3.3.2 Causal Models for Reinforcement Learning Agents

The intent in this work is not to provide explanations of *evidence* from the environment, but to provide explanations of the agent's behaviour based on the knowledge of how actions influence the environment. As such, we extend the notion of SCMs to include actions as part of the causal relationships.

We incorporate *action influence* models for MDP-based RL agents, extending SCMs with the addition of actions. We use the standard MDP notation described given in Chapter 2. The objective of an RL agent is to find a policy π that maps states to actions maximizing the expected discounted sum of rewards. We define the action influence model for RL agents as follows.

Formally, a signature S_a for an action influence model is a tuple $(\mathcal{U}, \mathcal{V}, \mathcal{R}, \mathcal{A})$, in which $\mathcal{U}$, $\mathcal{V}$,

and $\mathcal{R}$ are as in SCMs, and $\mathcal{A}$ is the set of actions.

Definition 3.3.2. An *action influence model* is a tuple $(S_a, \mathcal{F})$, where S_a is as above, and $\mathcal{F}$ is the set of structural equations, in which we have multiple for each $X \in \mathcal{V}$ — one for each *unique* action set that influences X. A function $F_{X.A}$, for $A \in \mathcal{A}$, defines the causal effect on X from applying action A. The set of *reward variables* $X_r \subseteq \mathcal{V}$ are defined by the set of nodes with an out-degree of 0; that is, the set of sink nodes. ■

We define the *actual instantiation* of a model M as $M_{\vec{\mathcal{V}} \leftarrow \vec{S}}$, in which $\vec{S}$ is the vector of state variable values from an MDP. In an actual instantiation, we set the values of all state variables in the model, effectively making the exogenous variables irrelevant.

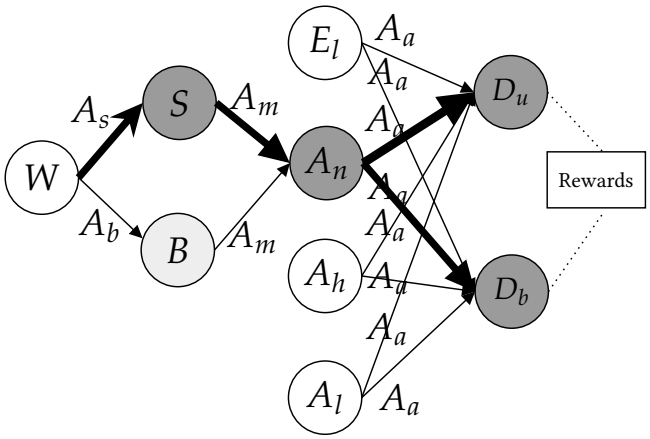
Figure 3.1 shows an action influence graph of the Starcraft II agent described in the previous section, with exogenous variables hidden. These *action influence models* are SCMs except that each edge is associated with an action. In the action influence model, each state variable has a *set* of structural equations: one for each *unique* incoming action. As an example, from Figure 3.1, variable A_n is causally influenced by S and B only when action A_m is executed, thus the structural equation $F_{A_n.A_m}(S, B)$ captures that relationship.

3.4 Explanation Generation

In this section, we present definitions that generate explanations from an action influence model. The process of explanation generation has 3 phases: 1) defining the qualitative causal relationships of variables as an action influence model; 2) learning the structural equations during RL; and 3) generating *explanans* from SCMs using the definitions given below.

We define an *explanation* as a pair that consist of: 1) an *explanandum*, the event to be explained; and 2) an *explanan*, the subset of causes given as the explanation [156]. Consider the example 'Why did you do P?' and the explanation 'Because of Q'. Here, the *explanandum* is P and *explanan* is Q. Identifying the *explanandum* from a question is not a trivial task. We define explanations for questions of the form 'Why A?' or 'Why not A?', where A is an action. In the context of a RL agent we define a *complete explanan* below.

Definition 3.4.1. A *complete explanan* for an action a under the actual instantiation $M_{\vec{\mathcal{V}} \leftarrow \vec{S}}$ is a tuple $\left(\vec{X}_r = \vec{x}_r, \vec{X}_h = \vec{x}_h, \vec{X}_i = \vec{x}_i\right)$, in which $\vec{X}_r$ is the vector of reward variables reached by



State variables:
W - Worker number
S - Supply depot number
B - barracks number
E - enemay location
A_n - Ally unit number
A_h - Ally unit health
A_l - Ally unit location
D_u - Destoryed units
D_b - Destroyed buildings
Actions:
A_s - build supply depot
A_b - build barracks
A_m - train offensive unit
A_a - attack

Figure 3.1: Action influence graph of a Starcraft II agent

following the causal chain of the graph to sink nodes; $\vec{X}_h$ the vector of variables of the head node of action a, $\vec{X}_i$ the vector of intermediate nodes between head and reward nodes, and $\vec{x}_r$, $\vec{x}_h$, $\vec{x}_i$ gives the values of these variables under $M_{\vec{\mathcal{V}} \leftarrow \vec{s}}$. ■

Informally, this defines a complete explanan for action a as the complete causal chain from action a to any future reward that it can receive. From Figure 3.1, the causal chain for action A_s is depicted in bold edges, and the extracted explanan tuple $([S = s], [A_n = a_n], [D_u = d_u, D_b = d_b])$ is shown as darkened nodes. We use depth-first search to traverse the graph until all the sink nodes are reached from the head node of the action edge.

3.4.1 'Why?' Questions

Lim, Dey, and Avrahami [135] found that the most demanded explanatory questions are *Why* and *Why not* questions. To this end, we focus on explanation generation for *why* and *why not* questions in this work.

Minimally Complete Explanations

Striking a balance between *complete* and *minimal* explanations depend on the epistemic state of the explainee [156]. We assume that we know nothing about the epistemic state of the explainee.

Recall from the definition of *explanans* (Definition 3.4.1), a 'complete' explanation would include *explanans* of all the intermediate nodes between the head and reward node of the causal chain. Clearly, for a large graph, this risks overwhelming the explainee. For this reason, we define *minimally complete* explanations.

McClure and Hilton [151] show that referring to the goal as being the most important for explaining actions. In our causal models, the rewards are the 'goals', but these alone do not form meaningful explanations because they are merely numbers. We define the human interpretable 'goal' using the variables in the predecessor nodes of the rewards. These define the immediate causes of the reward, and therefore which states will result in rewards.

However, rewards alone is only a longer-term motivation for taking an action. As such, we also include the head node of the action edge as the immediate reason for doing the action. We use this model to define our *minimally complete* explanations.

Definition 3.4.2. A *minimally complete* explanation is a tuple $\left(\vec{X}_r = \vec{x}_r, \vec{X}_h = \vec{x}_h, \vec{X}_p = \vec{x}_p\right)$, in which $\vec{X}_r = \vec{x}_r$ and $\vec{X}_h = \vec{x}_h$ do not change from Definition 3.4.1, and $\vec{X}_p = \vec{x}_p$ is the vector of variables that are immediate predecessors of any variable in X_r within the causal chain, with $\vec{x}_p$ the values in the actual instantiation. ■

Informally, for a complete causal chain, we take the first and last arcs of the causal chain, with their source and destination nodes, omitting intermediate nodes, as the minimal explanation. From Figure 3.1, for the action A_s, the minimally complete explanation is just the complete explanation, as there are no intermediate nodes.

Clearly, one could define other heuristics to decide which intermediate nodes to use as explanations, such as the knowledge of the explainee. However, in this chapter, we use this simple definition.

3.4.2 'Why not?' Questions

Why not questions let the explainee ask why an event has not occurred, thus allowing *counterfactuals* to be explained; something that is known to be a powerful explanation mechanism [156, 36]. Our model generates counterfactual explanations by comparing causal chains of the actual event occurred and the *explanandum* (counterfactual action). First, we define a *counterfactual instantiation* that specifies the optimal state variable values under which the counterfactual action B would be chosen.

Definition 3.4.3. A *counterfactual instantiation* for a counterfactual action B is a model $M_{\vec{Z} \leftarrow \vec{S_Z}}$, where $\vec{Z}$ gives the instantiation of all predecessor variables of action B with current state values *and* the instantiation of all successor nodes (of B) of the causal chain by forward simulating, using the structural equations. ■

Informally, this gives the 'optimal' conditions (according to the action influence model) under which we would select counterfactual action B, simulated through structural equations. We unravel this further in the Example 3.4.1 discussion using the Starcraft II scenario.

In the following definition, we use $\vec{X} = \vec{x}$ to represent the tuple $\left(\vec{X_p} = \vec{x_p}, \vec{X_h} = \vec{x_h}, \vec{X_r} = \vec{x_r}\right)$, and similar for $\vec{Y} = \vec{y}$ for readability.

Definition 3.4.4. Given a minimally complete explanation $\vec{X} = \vec{x}$ for action A under the actual instantation, and a minimally complete explanation $\vec{Y} = \vec{y}$ for action B under the counterfactual instantiation $M_{\vec{Z} \leftarrow \vec{S_Z}}$ (from Definition 3.4.3), we define a *minimally complete contrastive explanation* as the tuple $(\vec{X'} = \vec{x'}, \vec{Y'} = \vec{y'}, \vec{X_r} = \vec{x_r})$ such that $\vec{X'}$ is the maximal set of variables in $\vec{X}$ in which $(\vec{X'} = \vec{x'}) \cap (\vec{Y'} = \vec{y'}) \neq \emptyset$, where $\vec{x'}$ is then contrasted with $\vec{y'}$. That is, we only explain things that are different between the actual and counterfactual. This corresponds to the *difference condition* [155]. And $\vec{X_r}$ gives the reward nodes of action A. ■

Intuitively, a contrastive explanation extracts the actual causal chain for the taken action A, and the counterfactual causal chain for the B, and finds the differences.

Example 3.4.1. Consider the question, asking why a Starcraft II agent built supply depots, rather than choosing to build barracks:

Question Why not *build_barracks* (A_b)?

Explanation Because it is more desirable to do action build_supply_depot (A_s) to have more Supply Depots (S) as the goal is to have more Destroyed Units (D_u) and Destroyed buildings (D_b).

First we get the *actual instantiation* $m = [W = 12, S = 1, B = 2, A_n = 22, D_u = 10, D_b = 7]$ (instantiation should include all variables in the current state, only the required ones are shown for readability). The causal chain for the *actual* action 'why A_s?' would be as in Figure 1, and for the *counterfactual* action 'why not A_b?', the causal chain nodes would be $B \rightarrow A_n \rightarrow [D_u, D_b]$. We then get the *counterfactual instantiation* $m' = [W = 12, S = 3, B = 2, A_n = 22, D_u = 10, D_b = 7]$ using Definition 3.4.3. Applying the difference condition here, we obtain the minimally complete contrastive explanation (from Definition 3.4.4) as the tuple $([S = 1], [S = 3], [D_u = 10, D_b = 7])$ and contrast $[S = 1]$ with $[S = 3]$ to obtain the explanation of Example 3.4.1 (generated using a simple NLP template).

3.4.3 Learning Structural Causal Equations

Our approach so far relies on knowing the structural model, in particular, to determine the effects of counterfactual actions. *Why not* questions are inherently counterfactual [17], and having just the policy of an agent is not enough to generate explanations as counterfactuals refers to *possible* worlds that did not happen. Consider the Example 3.4.1, to generate this explanation, the optimal/maximum value of the state variable S is needed in the given time instance.

However, in model-free reinforcement learning, such environment dynamics are not known. Learning a complete model of the environment is a difficult problem. However, given a graph of causal relations between variables, learning a set of structural equations that are approximate yet 'good enough' to give counterfactual explanations may be feasible.

To this end, we assume that a DAG specifying causal direction between variables is given, and learn the structural equations as multivariate regression models during the training phase of the RL agent. We perform experience replay [160] by saving $e_t = (s_t, a_t, r_t, s_{t+1})$ at each time step t in a data set $D_t = \{e_1, ..., e_t\}$. Then we update the sub-set of structural equations $F_{X.A}$ using a regression learner $\widehat{\mathbb{L}}_{(s,a,r,s')\sim U(D)}$, in that we *only* update structural equations associated with the

specific action in the experience frame, drawn uniformly as mini-batches from D. For example, from Figure 3.1, for any experience frame with the action A_s, only the equation $F_{S.A_s}(W)$ will be updated. Any regression learner can be used as the learning model $\widehat{\mathbb{L}}$, such as a linear regressor or a multi-layer perceptron regressor.

While this approach may seem similar to learning environment dynamics of model-based RL methods, we only learn the structural equations, and we are only after an approximation that is good enough for explaining instances. Thus they can be approximate but still useful for explanation. Further, specifying the assumptions about the causal direction between variables is a much easier problem to encode by hand, and can be tested with the data.

3.5 Computational Evaluation

We evaluate *action influence models* in 5 OpenAI RL benchmark domains [31] and in the Starcraft II domain. The goal of this evaluation is to determine if learning action influence models leads to models that are faithful to the problem. Task prediction accuracy and training time for the structural causal equations are measured. The purpose of task prediction is to evaluate if the model is accurate enough to predict what an agent will do next, under the assumption that if it is not, then the model will not be of use to a human explainee.

We computationally simulate task prediction using Algorithm 1. Here we instantiate all the equations (which are the set of regression models $\mathcal{L}$) with the values of the current state S of the agent. We identify the equation that has maximum difference with the predicted state variable value and the actual, then get the action associated with it. This is informed by the reasoning that the agent will try to follow the optimal policy, and the action with the biggest impact to correct the policy will be executed. The impact is measured by the above mentioned difference. This is itself an approximation, but is a useful guide for task prediction.

We use linear SGD regression (LR), decision tree regression (DT) and multilayer perceptron regression (MLP) as the learners that approximate the structural equations. We choose benchmark domains based on varying levels of complexity, size (state features/number of actions) and train them using various RL algorithms to demonstrate the robustness of the model. Table 3.1 summarises the results of task prediction and time taken to train the structural equations given the replay data.

Algorithm 1 Task Prediction:Action Influence Model

Input: trained regression models $\mathcal{L}$, current state S_t
Output: predicted action a

$\vec{F_p} \leftarrow []$; vector of predicted difference
for every $\widehat{L} \in \mathcal{L}$ **do**
    $P_y \leftarrow \widehat{L} \cdot predict(S_{x.t})$; predict variable S_y at S_{t+1}
    $\vec{F_p} \leftarrow |S_y - P_y|$; difference with actual S_y value
end for
return $max\left(\vec{F_p}\right) \cdot getAction()$

Env - RL		Accuracy (%)			Performance (s)		
	Size	LR	DT	MLP	LR	DT	MLP
Cartpole-PG	4/2	83.8	81.6	86.0	0.007	0.018	0.03
MountainCar-DQN	3/3	69.7	57.8	69.6	0.020	0.037	0.32
Taxi-SARSA	4/6	68.2	74.2	67.9	0.001	0.001	0.49
LunarLander-DDQN	8/4	68.4	63.7	72.1	0.002	0.002	0.33
BipedalWalker-PPO	14/4	56.9	56.4	56.7	0.010	0.015	0.41
Starcraft-A2C	9/4	94.7	91.8	91.4	0.144	0.025	3.33

Table 3.1: Action influence model evaluation in 6 benchmark reinforcement learning domains (using different RL algorithms, PG, DQN etc.), measuring mean task prediction accuracy and training time of the structural causal equations in 100 episodes after training.

Overall, the results show the model did a reasonable job of task prediction, providing evidence that this could be useful for explanations. In domains that had discreet actions (all except the bi-pedal walker domain) the task prediction accuracy was close to 70% and over. This implies that computationally, action influence models are able to perform task prediction with only using structural causal equations accurately 70% of instances, where this accuracy can translate to a similar level of humans understanding of the agent when explanations are generated using the action influence models. Domains that have a clear causal structure (e.g Starcraft) performs best in task prediction. Starcraft II has a clear causal structure due its action set being highly dependant on the execution of previous actions which influences specific variables. In contrast domains like bi-pedal walker has actions that influence every variable (as continuous actions and features). Considering the performance cost it incurs, there was little gained by using MLP to approximate the equations, where in most cases linear regression is adequate. Apart from the BipedalWalker domain, our model performs well in task prediction with a negligible performance hit. The bipedalWalker domain has continuous actions, which our current model cannot handle

accurately. We plan to extend our model to continuous actions in future work.

3.6 Empirical Evaluation: Human Study

A human-grounded evaluation is essential to evaluate the explainability of a system, thus we carry out human-subject experiments involving explaining RL agents. We present two main hypotheses for the empirical evaluation; **H1)** Causal-model-based explanations build better mental models of the agent leading to a better **understanding** of its strategies (We make the assumption here that there is no intermediate effect on the mental model from other sources); and **H2)** Better understanding of an agent's strategies promotes **trust** in the agent.

We conducted a 4-condition between-subject study, with each condition having 2 within-subject variables. Here, the 4 independent conditions are given by the explanation generation models, while the 2 within-subject variables represent 'familiar' and 'novel' scenarios of the agent. Further details of these conditions are provided in the experiment parameters section below.

3.6.1 Methodology

We use StarCraft II, a real-time strategy game and a popular RL environment [235] as the domain. We implemented a RL agent for our experiment that competes in the default map.

To evaluate hypothesis (H1), we use the method of *task prediction* [110]. Task prediction can provide a quick view of the explainee's mental model formed through explanations, where the task is for the participant to predict 'What will the agent do next?'. We use the 5-point Likert *Explanation Satisfaction Scale* developed by Hoffman et al. [110, p.39] to measure the subjective quality of explanations. To evaluate hypothesis (H2), we use the 5-point Likert *Trust Scale* of Hoffman et al. [110, p.49]. We obtained ethics approval from The University of Melbourne human research ethics committee (ID-1953619.1).

Experiment Design

We use a recording of a full gameplay video (22 min) with the RL agents playing against in-game bot AI. The experiment has 4 phases.

Phase 1 involves collecting demographic information and training the participants. Using five gameplay video clips, the participant is trained to understand and differentiate the actions of the agent.

In phase 2, a clip of the gameplay video (15 sec) is played in a web-based UI, with a textual description of the scene. The participant can select the question type (why/why not) and the action, which together forms a question 'Why/Why not *action A*?'. Then, the textual explanation for the question with a figure of the relevant sub-graph of the agent's action influence graph is displayed. Explanations are pre-generated from our implemented algorithm. The participant can ask *multiple* questions in a single gameplay video. After every gameplay video, the participant completes the *Explanation Satisfaction Scale*. This process is repeated so we have data for each participant from five videos.

In Phase 3, we measure the *understanding* the explainee has after seeing the gameplay and the explanations. We measure understanding using the task prediction method as follows: the participant is presented with another gameplay video (10 sec), and presented with three selections of textual descriptions of what *action* the agent will do in *next* step; the participant selects an option, which includes 'I don't know'. We expect the participant is projecting forward the *local strategy* of the agent using their mental model. This mental model is formed through (or helped by) explanations seen in phase 1. This process is repeated for 8 tasks. In 4 of the task predictions, the behaviour is explainable using a causal chain previously seen in the training, but with different variable values. In the other 4 tasks, the behaviour is novel, but can be inferred by combining causal chains from different training tasks. In Phase 4, the participant completes a 5-point *Trust Scale*.

We conducted the experiments on *Amazon MTurk*, a crowd-sourcing platform popular for obtaining data for human-subject experiments [35]. The experiment was fully implemented in an interactive web-based environment. We excluded noisy data of users in 3 ways. First, we tested participants to ensure they had learnt about the agent's actions by prompting them to identify them. If the participant failed this, the experiment did not proceed. Then, for participants who completed, we omitted their data from analysis based on two criteria: 1) if the threshold of the time the participant spent on viewing explanations and answering tasks is below a few seconds, which was deemed too short to learn anything useful; and 2) if the participant's textual

responses to explain their task prediction choice were gibberish text or a 1-2 word response, as this indicated lack of engagement and care in the task. We controlled for language by only recruiting participants from the US.

Experiment Parameters

The experiment was run with 4 independent variables. We tested abstract (C) and detailed (D) versions of our action influence models and 2 baseline models described below: 1) Gameplay video without any explanations (N); 2) Relevant variable explanations (R). These explanations are generated using state relevant variables using *template 1* of Khan, Poupart, and Black [122, p.3] and visualized through a state-action graph, e.g 'Action *A* is likely to increase *relevant variable P*'; 3) Detailed action influence model explanations, where the causal graph is augmented to include atomic actions.

The above 4 independent variables (explanation models) are detailed below. Generated explanations of each model are given as examples.

- No-explanations (N): The video and a textual description of the agent behaviour is provided. The participant cannot query further about the behaviour (as there is no explanation model to query from). E.g: 'AI agent is doing the action *build_supply_depot*'.
- Relevant-variable explanations (R): Explanations are generated using template 1 of Khan, Poupart, and Black. The participant can query further about the agent behaviour using *why* questions (why not questions are not available as this explanation model cannot generate causal counterfactuals). A state-action figure that highlights the relevant variable is shown to the participant. E.g.: 'Because, the action *build_supply_depot* is likely to make the number of Supply Depots 1'
- Causal explanations (C): We use our action influence model and it's explanation generation methods. The used action influence model is abstract, in that the action in the model is concatenated (e.g. in StarCraft II, to build a supply depot, 3 separate granular actions need to be performed, which are 1. selecting a worker, 2. selecting a space in the map, and 3. commanding the worker to build. In the abstract model, we consider these 3 actions as one action - *build_supply_depot*) Participants can query the agent behaviour with both

why and why not questions. The action influence diagram is visualised and the relevant chain is shown to the participant. E.g.: 'Because, the goal is to increase Killed units and Buildings destroyed: Which depends on Army number: Army number is influenced by *build_supply_depot*'.

- Detailed causal explanations (D): We use the action influence model with the granular StarCraft II actions. The action influence diagram is visualised and the relevant chain (which is more detailed) is shown to the participant. E.g.: 'Because, the goal is to increase Killed units and Buildings destroyed: Which depends on Army number: Army number is influenced by *build_supply_depot*'. Note that although the explanation for this query is the same as the causal explanation (C), the presented action influence diagram is more detailed with granular actions.

We used power analysis to determine the needed sample sizes for the 4-conditions. Pilot runs were done for the conditions, having $n = 5$ participants for each condition. We consider the conditions causal explanations (C) and relevant-variable explanations R and calculate Cohen's d and obtain the effect size of 0.744. Using this effect size and $power = 0.8$ and significance level $\alpha = 0.05$, we do power analysis for both T-test and F-test (ANOVA) and obtain values 29.33 and 23.97 respectively. Thus we determined the total number of samples needed as $n = 120$.

We ran experiments for 120 participants, allocated evenly to the independent variables. Only 2 participants failed the test condition (learning of the agent's action, with repeated trials). Using the criteria described in the section Experiment Design, we excluded 18 participants from the analysis to yield 120 participants. The breakdown of the excluded data points are as follows, 1). Participants didn't exceed the time threshold - 4 (breakdown: conditions N - 1, R - 0, D - 2, C - 1), 2). Participants that had gibberish or very short text responses - 14 (breakdown: conditions N - 1, R - 4, D - 5, C - 4). Each experiment ran for approximately 40 minutes. We scored each participant on task prediction, 2 points for a correct prediction; 1 for responding 'I don't know' and 0 for an incorrect prediction for a total of 16 points. Scores were tallied. We compensated each participant with 8.5USD. Of the 120 participants, 36 were female, 82 male and 2 were not given. Participants were aged between 19 to 59 ($\mu = 34.2$) and had an average self-rated gaming experience and Starcraft II experience of 3.38 and 2.02 (5-point Likert) respectively.

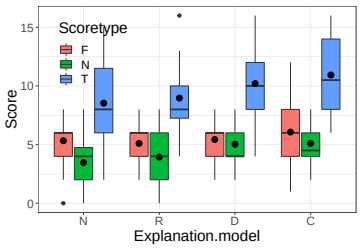


Figure 3.2: Box plot of task prediction scores of explanation models, T=total score, F=familiar score, N=novel score (higher is better, means represented as bold dots).

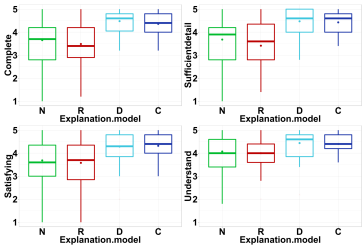


Figure 3.3: Box plot of explanation quality (likert scale 1-5, higher is better, means represented as dots).

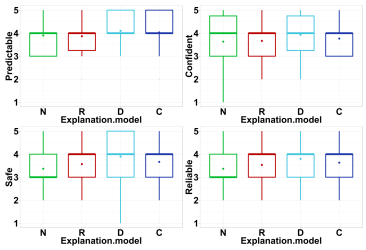


Figure 3.4: Box plot of trust (likert scale 1-5, higher is better, means represented as dots).

3.7 Results

Task Prediction

For the first hypothesis, the corresponding null and alternative hypotheses are: 1) $H_0 : \mu_C = \mu_R = \mu_D = \mu_N$; 2) $H_1 : \mu_C \geq \mu_R$; 3) $H_2 : \mu_C \geq \mu_D$; 4) $H_3 : \mu_C \geq \mu_N$, where abstract causal explanations (our model), detailed causal explanations, relevant variable explanations, and no explanations are given by C, D, R, and N respectively.

We conduct one-way ANOVA (Figure 3.2 illustrates the task score variance with explanation models). We obtained a p-value of 0.003 ($\mu_C = 10.90$, $\mu_D = 10.20$, $\mu_R = 8.97$, $\mu_N = 8.53$). Further values obtained by the one-way ANOVA test is as follows (sum of squares=109.8917, degrees of freedom=3, mean squares=36.6306, F-value=4.7684, eta-square effect size=0.1098). Thus we conclude there are significant differences between models on task prediction scores. We performed Tukey multiple pairwise-comparisons to obtain the significance between groups. From Table 3.3, the differences between the causal explanation model paired with other explanation models are significant for C-R and C-N pairs with p-values of 0.006 and 0.034. Additionally, we calculate the effect of the number of questions on the score, and obtain no statistical correlation using a correlation test (number of questions vs score, p = 0.33, model C) among same models. Because participants could select "I don't know" and receive 8 out of 16, we also further analyse scores based on 2 = 'correct', 0 = 'incorrect or 'I don't know', and obtain results that are still significant (p=0.004), means (C=10.90, D=10.10, R=8.93, N=8.47), for model pairs (C-N p=0.005, C-R p=0.035). We conducted Pearson's Chi-Square as a non-parametric test on the task prediction

scores, which showed significant results (p-value = 0.008, X-squared = 17.281).

Familiar and novel scenario type analysis: We further analyse the task prediction scores for familiar and novel scenarios (Depicted in Figure 3.2). For the familiar scenarios we perform one-way ANOVA and obtain values (F-value=1.426, p-value=0.238, effect-size=0.035) and for the novel scenarios we obtain values (F-value=5.023, p-value=0.002, effect-size=0.115). We conclude that the differences between the novel scenarios across explanation models are statistically significant.

Therefore we reject H_0 and H_2 and accept all other alternative hypotheses. Our results show that causal model explanations lead to a significantly better *understanding* of agent's strategies than the 2 baselines we evaluated, especially against previous models of relevant explanations. Participants did slightly worse on tasks with novel behaviour.

Effect of why and why not questions: To mimic real-world world human-agent explanations, participants were given the freedom to ask any number of questions (both why and why not) from the models that have the ability to generate explanations. Note that in the no explanation (N) condition, participants cannot ask questions as there is simply a description of the agent behaviour. Similarly, the relevant-variable (R) condition, only why questions can be asked due to the model's inability to simulate counterfactuals. Table 3.2 summarises the analysis of the effect of why and why not questions on the task prediction score. From the Table 3.2, we see that the total number of questions asked for each model were; C=340, D=328, R=224 and N=0. Breaking down the question number further; why questions asked were, C=190, D=182, R=224 and N=0; and why not questions asked were, C=150, D=146, R=0 and N=0. We further used a t-test to investigate whether there are any significant differences between the number of different types of questions asked in each condition. Note that we can only compare model pairs that can generate questions for the same type of questions (i. e. for why questions, models C, D and R; for why not questions models C and D). For why questions we obtain the following values, C-R model pair (statistic=-1.98, p-value=0.052), C-D model pair (statistic=0.48, p-value=0.63). Below values are obtained for the why not questions, C-D model pair (statistic=0.13, p-value=0.89). Though for why questions there is a barely significant difference (p-value = 0.52) between model pair C-R, for others, question number does not seem to differ significantly.

Question type	Model	Question-number	Mean	Median	Standard-deviation	Variance
Why	C	190	6.33	7	2.32	5.42
	D	182	6.06	6	1.86	3.46
	R	224	7.46	8	2.01	4.04
	N	0	0	0	0	0
Why not	C	150	5	4	3.91	15.33
	D	146	4.86	5	3.58	12.84
	R	0	0	0	0	0
	N	0	0	0	0	0
All	C	340	11.33	11	5.54	30.75
	D	328	10.93	10.5	4.87	23.79
	R	224	7.46	8	2.01	4.04
	N	0	0	0	0	0

Table 3.2: Effect of why and why not questions on the task prediction score. Explanation models are given by letters N, R, D, C.

Model pair	mean-diff	lwr	upr	p-value
C - N	**2.400**	0.534	4.265	**0.006**
C - R	**1.966**	0.101	3.832	**0.034**
D - N	1.666	-0.198	3.532	0.097
D - R	1.233	-0.632	3.098	0.316
C - D	**0.733**	-1.132	2.598	**0.735**
R - N	0.433	-1.432	2.298	0.930

Table 3.3: Pairwise-comparisons of explanation models of task prediction scores (higher positive diff is better)

Metric	Mdl-pair	Mean-dif	Median-dif	p-val
Complete	C-N	0.707	0.700	0.061
	C-R	0.873	1.000	**0.012**
Sufficient	C-N	0.746	0.700	**0.039**
	C-R	1.013	1.000	**0.002**
Satisfying	C-N	0.633	0.800	0.082
	C-R	0.740	0.700	**0.029**
Understand	C-N	0.326	0.400	0.497
	C-R	0.400	0.400	0.316

Table 3.4: Explanation quality (likert scale data 1-5)

Explanation Quality

To measure the subjective quality of the explanations, we used the explanation quality survey of Hoffman et al. [110, p.39]. The survey includes 8 questions having a 5-point Likert scale. Though we presented the whole survey to the participants, we only report results on the first 4 questions.

The last 4 questions measure 'how to use the software', 'useful for the user's goal', 'accuracy' and 'trust and not trust'. These last 4 questions are not relevant to the reinforcement learning agent we developed because of its strategy-oriented nature. Further, the last question is omitted because trust is later analysed with a different survey.

Figure 3.3 depicts the likert scale data of explanation metrics (understand, satisfying, sufficient detail and complete) for aggregated video explanations of explanation models. As before we performed a pair-wise ANOVA test, results are summarised in Table 3.4. Our model obtained statistically significant results and outperformed the benchmark 'relevant explanation' (R) for all metrics except 'Understand'.

Trust

We use the Trust survey of Hoffman et al. [110, p.49] to measure the subjective user trust towards the agent. This survey consists of 8 questions with questions having a 5-point Likert scale. Similar to the explanation quality survey, we present the whole survey to the participants. The last 4 questions of the survey; 'efficient', 'wary of the tool', 'better task performance', 'use for decision making' are not relevant for our agent setting. Thus we report results on the first 4 questions of the survey.

For the second main hypothesis (H2) that investigate whether explanation models promote trust, the obtained p-values for trust metrics *confident*, *predictable*, *reliable* and *safe* were not statistically significant (using pair-wise ANOVA). Although the difference is not significant we can see causal models have high means and medians (see Figure 3.4). We conclude that while the explanation quality and scores are significantly better for our model, to promote trust further interaction is necessary; or perhaps our RL agent is simply not a trustworthy Starcraft II player.

We answered the main hypothesis (H1) based on the task prediction scores and the explanation quality results. While the trust results of the hypothesis H2 does not vary significantly across different models, from higher means and medians, causal model based explanations seems to invoke higher level of trust. This has implications on how the explanation model can effect the generation of trust, when the interaction between the explainee and the explainer spans a longer time frame.

We further analysed self-reported demographic data to see if there is a correlation between task prediction scores and self-reported Starcraft II experience level (5-point Likert). Pearson's correlation test was not significant (p=0.45) thus we conclude there is no correlation between scores and experience level. This can possibly be attributed to our Starcraft II scenario differing from the standard game.

A limitation of our experiment is that we made a strong linearity assumption for Starcraft II, which enabled linear regression to learn SCMs for a relatively small number (9) of state variables.

3.8 Conclusion

In this Chapter, we introduced *action influence* models for model-free reinforcement learning agents. Our approach learns a structural causal model (SCM) during reinforcement learning and has the ability to generate explanations for *why* and *why not* questions by counterfactual analysis of the learned SCM. We computationally evaluated our model in 6 benchmark RL domains on *task prediction*. We then conducted a human study (n=120) to evaluate our model on 1) task prediction, 2) explanation 'goodness' and 3) trust. Results show that our model performs significantly better in the first 2 evaluation criteria. One weakness of our approach is that the causal model must be given beforehand. Chapter 5 will focus on learning the the action influence structure using the agent's state-action interaction traces.

Chapter 4

Distal Explanations for Reinforcement Learning Agents[1]

4.1 Introduction

Explanation models that emulate human models of explanations have the potential to provide intuitive and natural explanations, allowing the human a deeper understanding of the agent [58, 1, 156, 241]. There exists a large body of literature in cognitive psychology that studies the nature of explanations. One prevalent theory is that explanations are innately *causal* [92]. Causal explanations resonate with humans as we make use of *causal* models of the world to encode cause-effect relationships in our mind [210], and leverage these models to explain why *events* happen. Causal models also enable the generation of *counterfactual* explanations—explanations about events that did not happen but could have under different circumstances [92]. So causal explanations have the potential to provide 'better' explanations to humans.

Recent work in the XAI research community has demonstrated the effectiveness of causality and causal explanations for interpretability and explainability [36, 125, 89, 203]. In the context of model-free reinforcement learning (RL) agents, causal models have been encoded using *action influence graphs* to generate explanations using *causal chains* and in the work presented in Chapter 3 show to support subjectively 'better' explanations and yield improved performance in *task prediction* [110] as compared with state-action based explanations [122]. While action influence models provide a skeleton to generate causal explanations for RL agents, finer details

[1]This chapter is adapted from the pre-print article under review: "Distal explanations for explainable reinforcement learning agents." arXiv preprint arXiv:2001.10284 (2020).

of the composition of causal explanations can be absent. We argue that, through investigating interactions of RL agents and humans, some shortcomings of action influence models can potentially be alleviated.

To ground the effect that explanation models have on human explanation, we conduct human-agent experiments on how humans formulate explanations of agent behaviour. Participants of this study received explanations from three different models: visual agent behaviour explanations; state-action based explanations [122]; and causal explanations. Then participants were asked to formulate their own explanations of the agents' behaviour as a textual input. The study was carried out with 30 participants and we obtained 240 explanations in total. We used thematic analysis [30] to identify recurring concepts present in the explanations.

Results of our analysis show that while causality was indeed present, these self-provided explanations predominantly referred to a *future* action that was dependent on the current action. Participants' tendency to include a future action in their explanations indicates an understanding of the causal chain of actions and events. This phenomenon is well explored in cognitive psychology and is defined as *opportunity chains* [109]. We use insights gained from the human-agent study to inform our design of an explanation model that can explain opportunity chains and the future action termed the *distal* action.

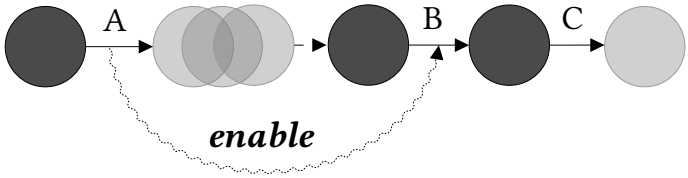


Figure 4.1: An *opportunity chain* [109], where event A *enables* B and B causes C.

Hilton, McClure, and Slugoski [109, 152, 108] note that humans make use of *opportunity chains* to describe events through causal explanation. An opportunity chain takes the form of A enables B and B causes C (depicted in Figure 4.1), in which we call B the 'distal' event or action. Simply, the distal action is the action that is being *enabled* in the future by some action that occurs before it. If A fails to occur, the distal action B cannot be executed. For example, an accident can be caused by slipping on ice which was *enabled* by water from a storm the day before. Here, the distal action is the 'slipping on ice' which was enabled by 'rain water'. Opportunity chains are *causal chains* that can be extracted from action influence models. Thus action influence models

can be used as a platform to augment causal explanations with opportunity chains.

To that end, we propose a *distal* explanation model that can generate opportunity chains as explanations for model-free RL agents. We provide definitions for distal explanations and learn the opportunity chains of extracted causal chains using a recurrent neural network [202]. A distal explanation by itself would not make a *complete* explanation. For this reason, we use action influence models to get the agent's 'goals'. We further improve upon action influence models by using decision trees to represent the agent's policy.

We computationally evaluate the accuracy of task prediction [110, p.12] and counterfactuals in 6 RL benchmark domains using 6 different RL algorithms, and show that our distal explanation model is robust and accurate across different environments and algorithms. Then we conduct human experiments using RL agents trained to solve 3 different scenarios, where agents solve 1) an adversarial task; 2) a search and rescue task; and 3) a human-AI collaborative build task. The human study was run with **90** participants, where we evaluate task prediction [110] and explanation satisfaction. Results indicate that our model performs better than the two tested baselines.

The main contribution in this chapter is twofold: 1) we introduce a distal action explanation model that is grounded on human data; 2) we extend action influence models by using decision trees to represent the agent's policy and formalise explanation generation from decision nodes and causal chains. As secondary contributions, we also provide the coded corpus of human-agent experiment with **240** explanations and two custom maps that are suited for explainability in the StarCraft II environment.

4.2 Related work

In this section we discuss the body of literature that explores explainability in a human-centered manner. Literature relating to explainable reinforcement learning is discussed in Chapter 2.2.

4.2.1 Human-Centred Explanation

Some researchers have recently emphasised how humans models of explanations can benefit XAI systems [156] and how humans expect familiar models of explanations from XAI systems [58].

Though some recent progress has been made, human-centred computational models is still in its infancy.

Hilton, McClure, and Slugoski [109, 152, 108] has explored how causal chains of events inform and influence the explanations of humans. *Opportunity chains* can inform the explainee about long term dependencies that events have on each other, where certain events *enable* others. Human experiments have also been carried out that investigate he effects of opportunity chains on human-to-human explanation [109]. However, this work has not yet been extended to the case of model-free MDPs.

Our proposed *distal explanation* model takes insights from social psychology literature to combine *opportunity chains* with *causal* explanations. To the best of our knowledge this is the first of such model in the context of explainable reinforcement learning agents.

4.3 Human-Agent Study: Insights from Human Explanations

In this section, we discuss insights we can gain from human models of explanation in literature. We then ground these models in data by conducting a human-agent experiment.

4.3.1 Human Models of Causal Explanation

Causality is a recurring concept in explanation models of social psychology and cognitive science literature [106, 111, 142]. Using causal models as the basis for explanation seems natural and intuitive to humans [210], since we build causal models to represent the world and to reason about it. Thus, it is plausible that, when used in intelligent agents, causal models have the ability to provide 'good' explanations to humans.

Importantly, causal models consist of *causal chains*. A causal chain is a path that connects a set of events, where a path from event A to event B indicates that A has to occur before B [156] (we use event and action interchangeably in the chapter). Hilton, McClure, and Slugoski [109] define five types of causal chains that lead to five different types of explanations. Hilton, McClure, and Slugoski categorise these as, temporal, coincidental, unfolding, opportunity chains and pre-emptive. Through human experiments, Nagel and Stephan [165] demonstrated that *distal* causes forms significant portion of an explainee's understanding of a terminal cause. Böhm and

Pfister [27] also affirms that, humans give both proximal and distal causes as explanations. Its important to note that, while in cognitive psychology literature a distal cause is a remote cause of an event in the past (essentially looking 'backward' from an event), in our agent simulations we use the distal terminology to denote an 'action' that is remote in the future (according to the agent's viewpoint this is looking 'forward' from a present event/action). Hilton, McClure, and Sutton [108] also explored how humans select different causal chains to provide explanations through human experiments. We conduct a similar study to gain insights from human models of explanation in a human-agent setting, and report results below.

4.3.2 Study Objectives

We seek to investigate how humans provide explanations of intelligent agents' behaviour and what concepts are present in such explanations. In contrast to similar studies done in social psychology [108], our experiments present explanations of the agent's behaviour first to the participant and then gives the freedom to form their own explanations of the agent. The main objective of the study is to discover the frequency of different concepts in these human generated explanations given the agent behaviour explanations using different explanation methods.

4.3.3 Experiment Design

We conducted a human-agent study with 30 participants. In the first phase, participants were shown reinforcement learning agents playing the game StarCraft II. The agent behaviour (policy) was explained by providing 'local' explanations of agents' actions using one of 3 different explanations models: 1) No explanations, just visual description of the agent's behaviour; 2) State-action based explanations [122]; and 3) Causal explanations. Participants were divided evenly for each of these explanation models. Experiment was run on a web based interactive interface in through the *Amazon Mechanical Turk* [35].

In the second phase, participants were shown new agent behaviour and were asked to 'predict' the agent's next action. Participants are expected to predict the next action based on the learned model of the agent in the first phase through explanations. This prediction task is not important to the objectives of this study, but is used as a way to get the participants to reasong about behaviour. In the same page, participants were then asked formulate their own explanations

Table 4.1: Codes (of the concepts) and descriptions of human generated explanations of agent behaviour. Examples are given from different participants.

Code	Description	Example
Action	An **action** of the agent	P10: "It will keep **attacking** while it has the advantage"
Feature	A **feature** of the agent	P4: "The optimal number of **supply depots** is 2. and they should build those before a **barracks**"
Temporal	Refer some **temporal** quality	P12: "I think the artificial player will want to train marines right away so that it has an army quickly and **be able to attack** the enemy."
Objective	Refer to a short term **objective**	P12: "I think the artificial player **will want to** train marines right away so that it has an army quickly and be able to attack the enemy."
Causality	Implies a **causal** relationship	P3: "**you need an army to attack**[action] and by training marines you can do that"
Quantitative	Refers to a **numerical** value of a **feature**	P4: "The optimal number of supply depots is **2**. and they should build those before a barracks"
Qualitative	A **qualitative** reference to a **feature**	P4: "**As long as they have enough** healthy marines. they should keep attacking"
Uncertainty	Mentions **uncertainty**	P5: "The army is in good health. **it will most likely** continue to attack."
Contrastive	**Contrasting** a **feature** with another	P15 "There are 2 **supply depots but only 1 barrack.**"
Goal	Refer to the **goal(s)** of the agent	P10: "The point of the game is **to kill the enemy** and **destroy their base**. so (incorrectly) the AI thinks the next step is to attack."

about the agent. Participants were given a text-box to input the formulated explanations with no restrictions to word limit. This process is repeated for 8 rounds.

To filter out devious participants, we used the following approaches. Explanations containing less than three words or gibberish text were omitted. We also considered the time it took to input the explanation as a threshold. We omitted six participants according to the above criteria. In total we obtained a total of 240 explanations.

4.3.4 Method

We use thematic analysis [30] to *code* the data and to identify concepts. By using thematic analysis, meaningful insights can be gained on how explanations of agents behaviour relates

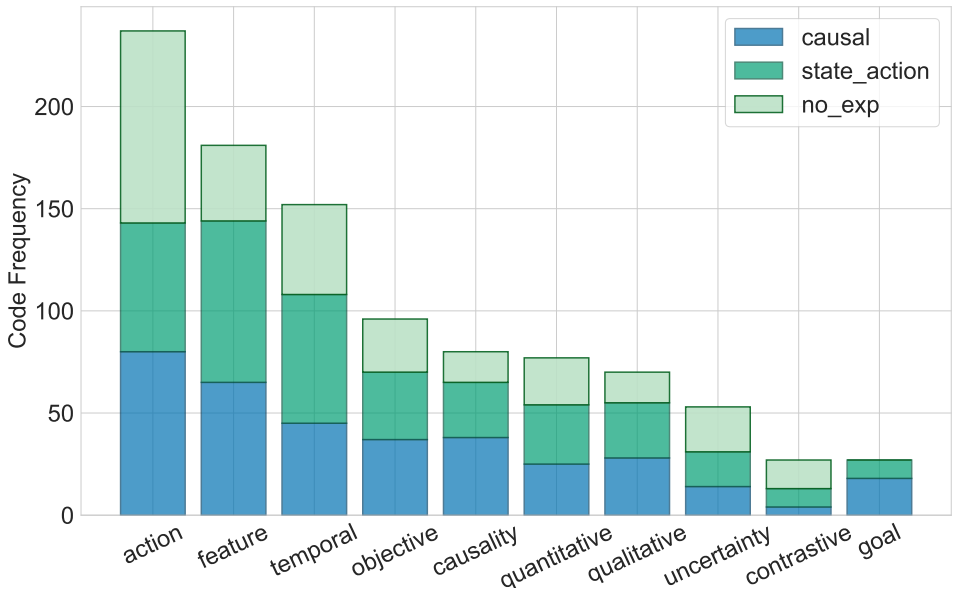


Figure 4.2: Codes and their frequencies of 240 human explanations of reinforcement learning agents (that were using 3 different explanation models)

to existing literature on human explanations. As the first step in the thematic analysis, each explanation will be be divided into small chunks to identify categories and then these will be divided further into codes. Intuitively, a 'code' represents an atomic concept that exist in the explanation corpus. For an example, when a reference to an 'action' of the agent is present in the explanation, the sub-string of that reference can be *coded* (tagged) as an **Action**. This process is done manually until all the data chunks and explanations are coded. To ensure correctness, further passes through the explanation corpus is done as an attempt to identify new concepts that might have been missed in the first pass. Coded concepts and their descriptions are given in Table 4.1, along with example explanations extracted from participants.

4.3.5 Results

Figure 4.2 shows the frequencies of 9 codes across the 3 explanation models of the RL agents. Participants referred to 'actions' and 'features' of the agent the most, and often included the 'objective' or the 'goal' of the agent, which is present in action influence models. Most importantly, the third most frequent code is 'temporal', in which participants refer to future actions the agent

will take (i.e. distal actions). For example, consider an explanation from the data corpus, "The AI will want to have barracks so that it can then train soldiers to engage in attacks. It will want to progress". Here, the participant's explanation contains the distal action 'train soldiers' which is *enabled* by 'have barracks'. 'Causality' is also present in the explanations, interestingly even in 'No explanation' and State-action based explanation models. This suggests that humans frequently associate causal relationships when generating explanations. Our human-agent experimental data reaffirm the presence of opportunity chains in causal chains [109], and show that these are frequently used to express how future actions are dependent on current actions of agents.

4.3.6 Discussion

Table 4.2: Presence of the concepts that were derived from codes, in different explainable reinforcement learning methods.

XRL Method	Action	Feature	Temporal	Causal	Contrast	Objec	Goal	Quan	Qual	Uncer
[69]	✓	✓						✓		
[122]	✓	✓			✓			✓		✓
[243]		✓				✓		✓	✓	✓
[236]	✓	✓			✓			✓		
[215]	✓	✓						✓		
[99]	✓	✓			✓	✓		✓		
[9]	✓	✓				✓		✓		
[224]		✓				✓	✓	✓		
[77]		✓						✓		
[117]		✓				✓	✓	✓		
Action Influence Models	✓	✓		✓	✓	✓	✓	✓	✓	
Proposed method	✓	✓	✓	✓	✓	✓	✓	✓	✓	

The concepts that were derived from the codes are present in previous explainable reinforcement learning methods to varying degrees. Table 4.2 shows how these concepts are distributed. As most of these methods were not developed in a ground-up manner, some important concepts present in human explanations were not implemented in their explanation generation. When developing novel explanation models, insights gained from human-agent studies can help ground the model in the characteristics of *human explanation*. Human grounded explainable models can be more effective and accepted when deployed [156, 131]. To this end, we conducted human-

agent experiments to discover how a human would explain the reasoning and behaviour of an agent, when the agent has given prior explanations of it's own actions. When these human explanations were abstracted into 'codes', notable concepts like 'causality' and 'temporality' emerged. Previous work done in social psychology support our findings and coincide well with notions like opportunity chains [109].

Though previous studies have explored the structure of causal chains in human explanations [108], these are largely done in the absence of an intelligent agent. Further, in [108], an explanation structure is investigated for events that have already occurred. In our study, as human explanations are for the *behaviour* of the agent, they can refer to how the past and present actions of the agent can influence the future. Ultimately, we use the resultant concepts of *causality* and *distal opportunity chains* to propose the distal explanation model for reinforcement learning agents.

4.4 Preliminaries

In this section, we present the necessary background that is required to follow the remainder of the chapter.

4.4.1 Markov Decision Processes

We concern ourselves with providing an explanations for Markov Decision Process (MDP) based model-free RL agents. An MDP is a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ give state and action spaces respectively (here we assume the state and action space is finite and state features are described by a set of variables ϕ); $\mathcal{T} = \{P_{sa}\}$ gives a set of state transition functions where P_{sa} denotes state transition distribution of taking action a in state s; $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a reward function and $\gamma = [0, 1)$ gives a discount factor. The objective of a reinforcement learning agent is to find a policy π that maps states to actions maximizing the expected discounted sum of rewards. In model-free reinforcement learning, $\mathcal{T}$ and $\mathcal{R}$ is not known and the agent does not explicitly learn them.

4.4.2 Structural Causal Models

Structural causal models (SCMs) [92] provide a formalism for representing variables and *causal* relationships between those variables. SCMs represent the world using random variables, divided into exogenous (external) and endogenous (internal), some of which might have causal relationships which each other. These relationships can be described with a set of *structural equations*. A detailed discussion of SCMs is done in Chapter 2 and 3.

Definition 4.4.1. A *signature* $\mathcal{S}$ is a tuple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$, where $\mathcal{U}$ is the set of exogenous variables, $\mathcal{V}$ the set of endogenous variables, and $\mathcal{R}$ is a function that denotes the range of values for every variable $\mathcal{Y} \in \mathcal{U} \cup \mathcal{V}$. ■

A *context* $\vec{u}$ is a vector of unique values of each exogenous variable $u \in \mathcal{U}$. A *situation* is defined as a model/context pair $(M, \vec{u})$. Given a situation $(M, \vec{u})$ an *instantiation* of M given $\vec{u}$ is defined by assigning all endogenous variables the values corresponding to those defined by their structural equations.

An *actual cause* of an event φ is a vector of endogenous variables and their values such that there is some counterfactual context in which the variables in the cause are different and the event φ does not occur. An explanation is those causes that an explainee does not already know. Following example gives perspective to the notions discussed above.

Example 4.4.1. Consider the *coffee task* [29] where a robot has to deliver coffee to a user. The state consists of six binary variables, robot location (L), robot is wet (W), robot has umbrella (Umb), raining (Rn), robot has coffee (C) and user has coffee (Usr). Actions of the robot are *go*, *buy coffee*, *get umbrella* and *deliver coffee*. Then we can identify the set of *endogenous* variables $\mathcal{U}$ as L, W, Umb, C and Usr because the values of these variables can be influenced by the actions of the robot. In contrast, variable the variable Rn (raining) is an *exogenous* ($\mathcal{V}$) variable, because it is not defined by a function. A *signature* for this is generated by combining $\mathcal{U}$, $\mathcal{V}$ and the value range the variables can take (in this case either 0 or 1). Having the signature at hand, we can formulate a *structural causal model* M by identifying the set of functions $\mathcal{F}$ that describe causal relationships of state variables. Assuming there is only one such function, we can define it $F_{Usr} = C + L$. This implies that the variable 'user has coffee' is causally influenced by variables 'robot has coffee' and 'robot location'. Model M can be *instantiated* by getting the current values

of the state variables and applying them to the set of $\mathcal{F}$. The *actual cause* of the event Usr being true is the vector $(C = 1, L = 1)$ as both of these variables needs to be true for the user to have the coffee.

For a more complete review of SCM's we direct the reader to [92].

4.4.3 Action Influence Models

Action influence models provide explanations of the agent's behaviour based on the knowledge of how actions influence the environment. Informally, action influence models are an extension of SCMs that are augmented with agent actions. These models capture the causal relationships that exist in agent's knowledge about the world (i.e. state variables). Action influence models are formally defined for RL agents as follows,

Definition 4.4.2. The *actual instantiation* of an action influence graph is defined as $M_{\vec{\mathcal{V}} \leftarrow \vec{S}}$, in which $\vec{S}$ is the vector of state variable values from an MDP and $\mathcal{V}$ as in Definition 3.3.1. A *counterfactual instantiation* for a counterfactual action B is a model $M_{\vec{Z} \leftarrow \vec{S_Z}}$, where $\vec{Z}$ gives the instantiation of a counterfactual state $\vec{S_Z}$. ■

In an *actual instantiation*, we set the values of all state variables in the model, effectively making the exogenous variables irrelevant. Similarly, a *counterfactual instantiation* assign values to the model M that could have realised under the action B.

Figure 3.1 shows the graphical representation of Definition 3.3.2 as an action influence graph of the StarCraft II agent described in the previous section, with exogenous variables hidden. These *action influence models* are SCMs except that each edge is associated with an action. In the action influence model, each state variable has a *set* of structural equations: one for each *unique* incoming action. As an example, from Figure 3.1, variable $\hat{A}_n$ is causally influenced by $\hat{S}$ and $\hat{B}$ only when action A_m is executed, thus the structural equation $\mathcal{F}_{A_n.A_m}(S, B)$ captures that relationship.

4.4.4 Explanations

An explanation is generally defined as a pair that contains; 1) an *explanandum*, the event to be explained and 2) an *explanan*, the subset of causes that explain that event [156]. In its simplest

form, the explanation for the question 'Why P?' would be in the form of 'Because Q'. In the above example, P is the explanandum and Q is the explanan. As Lim, Dey, and Avrahami (2009) notes, *why* and *why not* questions are the most demanded explanatory questions. In the context of RL agents, we are interested in answering 'Why A?' and 'Why not A?' questions. Here, A is an action of the agent and the explanation will be *local*.

Action influence models can be used to generate *minimally complete* explanations. An explanation that constitutes all the causes as an explanan risk overwhelming the explainee, thus it is important to balance the *completeness* and the *minimality* of the explanations [156].

Definition 4.4.3. A *minimally complete* explanation for an action a under the actual instantiation $M_{\vec{\mathcal{V}} \leftarrow \vec{S}}$ is a tuple $\left(\vec{R} = \vec{r}, \vec{H} = \vec{h}, \vec{I} = \vec{i}\right)$, in which $\vec{R}$ is the vector of reward variables reached by following the causal chain of the graph to sink nodes; $\vec{H}$ the vector of variables of the head node of action a, $\vec{I}$ is the vector of variables that are immediate predecessors of any variable in $\vec{R}$ within the causal chain, with $\vec{r}$, $\vec{h}$, $\vec{i}$ giving the values of these variables under $M_{\vec{\mathcal{V}} \leftarrow \vec{S}}$ from Definition 4.4.2. ■

McClure and Hilton [151] argue that 'goals' should be referred to in some form when explaining actions. In reinforcement learning, rewards of the agent can be thought of as a proxy for the goals. Though in most cases the 'rewards' ($\vec{R}$ from Definition 4.4.3) on itself would not form a complete explanation, because they are not attached to variables. Immediate predecessor nodes ($\vec{I}$) of the reward nodes refer to the state variables that 'trigger' rewards.. Though this combination now can explain the long term motivation of the agent, the head node ($\vec{H}$) attached to the action is used to explain the immediate (short-term) cause. From Figure 3.1, the explanation for *Why action* A_s would constitute, D_u and D_b in as reward variables $\vec{R}$, A_n in $\vec{I}$ and S in $\vec{H}$. Chapter 3 present a method for generating such explanations, and evaluate this on a large-scale user study.

4.5 Distal Explanation Model

From the insights gained from human explanations discussed in Section 4.3 we propose a distal explanation model that can generate explanations for opportunity chains. In the following sections we use the adversarial scenario (discussed at length in Section 4.6.1) of the StarCraft II environment as a running example to aid the definitions.

4.5.1 Overview

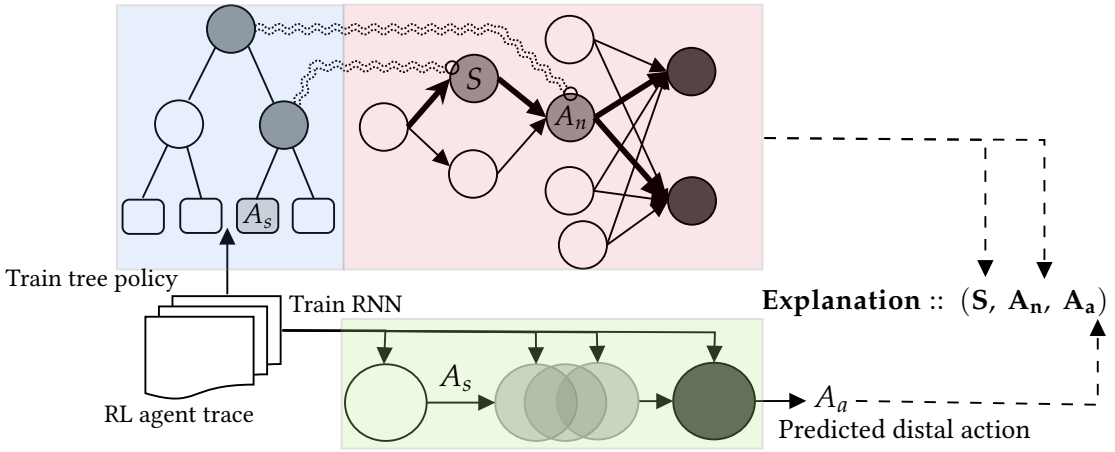


Figure 4.3: An overview of the Distal explanation model

Figure 4.3 shows an overview of the distal explanation model. The model consists of four distinct components. First, state-action pairs are extracted as a replay dataset from the episodes during reinforcement learning. Dataset generation happens at the agent training time. This dataset is used to train the decision-tree policy (indicated as the blue sub-component in Figure 4.3). The decision-tree policy is used as a surrogate policy for the agent, where it is used to extract reasons (in the form of decision nodes) for a given action (we discuss this process at length in Section 5.2). The dataset is also used to train the distal action predictor (shown as the green sub-component), which predicts dependent actions. Because we want to predict distal actions (contained in a opportunity chain) using a sequence of prior actions from the agent action trace, a many-to-one recurrent neural network [202] is used as the predictor, though other sequence predictors can also be used. An action influence graph is used to extract causal chains (shown in red) that is used in conjunction with the decision tree policy to produce the final explanation. The explanation is given as a three-tuple: reward nodes of the causal chain; matched decision nodes; and the predicted distal action.

Before formalising the distal explanation model we first discuss how explanations can be generated using decision tree policies.

4.5.2 Causal Explanations from Decision Trees

Although causal explanations from action influence models have been shown to perform better than state-action based [122] explanation models, the use of *structural equations* models the environment rather than the policy of the agent. Thus, the explanations from these model why an action would be a good idea, rather than why the agent chose it. In this work, we instead propose to extract reasons for action selection from a *surrogate* policy. We learn an interpretable surrogate policy in the form of a decision tree using batched replay data. If the agent's underlying policy is also a decision tree, this step can be omitted.

Training The Surrogate Policy: The *distal explanation* model we introduce uses decision nodes of a decision tree that represent a surrogate policy to generate explanations with the aid of causal chains from an action influence model. Let $\widehat{\mathbb{T}}$ be a decision tree model. In each episode at the training of the RL agent, we perform experience replay [136] by saving $e_t = (s_t, a_t)$ at each time step t in a data set $D_t = \{e_1, ..., e_t\}$. Drawing uniformly from D as mini-batches, we train $\widehat{\mathbb{T}}$ using input $x = \vec{s}$ and output $y = \vec{a}$. Clearly, explanations generated from an unconstrained decision tree can overwhelm the explainee, as these produce a large number of decision nodes for a question. Thus we limit the growth of $\widehat{\mathbb{T}}$ by setting the max number of leaves to the number of actions in the domain (i.e. the leaves of the trained $\widehat{\mathbb{T}}$ will be the set of actions of the agent). We later show that this hardly affects the task prediction accuracy compared to a depth unconstrained decision tree for our experiments. To get the decision nodes of $\widehat{\mathbb{T}}$ in state S_t, we simply traverse the tree from the root node until we reach a leaf node and get the nodes of the path. The decision tree of the StarCraft II adversarial task is given in Figure 4.4 a), with the decision nodes A_n and B for the action A_s. Each decision node maps to a feature variable of the agent's state. Figure 4.4 shows how the decision nodes are mapped to the action influence graph, in the StarCraft II adversarial scenario.

Generating Explanations Using the Surrogate Policy: In the context of an RL agent, we introduce a new definition of *minimally complete* explanations using decision nodes for 'why' questions below.

A primitive explanation can be generated by using the decision-tree policy alone, by extracting

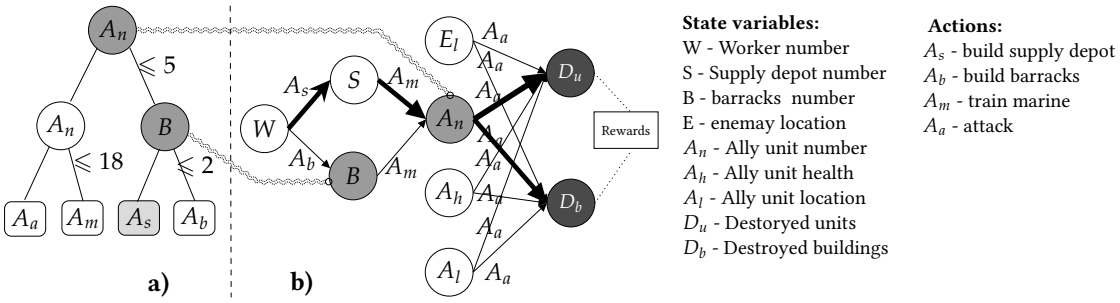


Figure 4.4: Generating explanations by mapping (a) decision nodes to (b) causal chains.

the decision nodes of an action. E.g. for the question *Why* A_s, we can obtain the decision nodes simply by traversing to the leaf node A_s from the root node (A_n and B are the decision nodes in this case, as highlighted in Figure 4.4 a)). However, an explanation like this can contain variables that are not *causally* relevant to the action performed. This primitive explanation can be enhanced by taking the causal chain for the action being explained from an action influence model and filtering out causally irrelevant variables. We define this as a minimally complete explanation below.

Definition 4.5.1. Given the set of decision nodes $\vec{X}_d = \vec{x}_d$ for the action a from a decision tree $\widehat{\mathbb{T}}$, we define a *minimally complete* explanation for a *why* question as a pair $\left(\vec{R} = \vec{r}, \vec{N} = \vec{n}\right)$, in which $\vec{R}$ is the vector of reward variables reached by following the causal chain of the graph to sink nodes; $\vec{N}$ is such that $\vec{N}$ is the maximal set of variables in which $\vec{N} = (\vec{X}_a = \vec{x}_a) \cap (\vec{X}_d = \vec{x}_d)$, where $\vec{X}_a$ is the set of intermediate nodes of the causal chain of action a, with $\vec{r}$, $\vec{x}_a$ and $\vec{x}_d$ giving the values under the actual instantiation $M_{\vec{\mathcal{V}} \leftarrow \vec{S}}$ from Definition 4.4.2. ■

Above definition only select the decision nodes (from the total set of decision nodes given from the decision-tree policy) that exist as intermediate nodes of the causal chain of the given action.

In the StarCraft II scenario, for the question 'Why action A_s?', we can generate the minimally complete explanation by first finding the decision nodes for action A_s, shown as medium grey nodes in Figure 4.4(a). Then finding the causal chain of action A_s (given by the bold path in Figure 4.4). And finally getting the common set of nodes from the causal chain and the decision nodes (B in Figure 4.4) and appending the reward nodes (D_u and D_b). Example 4.5.1 below compare and contrast an explanation with and without the use of action influence models.

Example 4.5.1. Question: Why A_s?

Algorithm 2 Generating Counterfactuals

Input: causal model $\mathcal{M}$, current state S_t, trained decision tree $\widehat{\mathbb{T}}$, *actual* action a, Δ
Output: contrastive explanation t $\vec{X_c}$

1: $\vec{X_d} \leftarrow \widehat{\mathbb{T}} \cdot traversetree(a)$;vector of decision nodes of a from $\widehat{\mathbb{T}}$
2: $\vec{X_c} \leftarrow []$;vector of counterfactual decision nodes.
3: **for** every $D \in \vec{X_d}$ **do**
4: $x_d \leftarrow D \cdot decisionNodeValue()$; decision boundary value of D
5: $x_m \leftarrow D \cdot moveBoundary(x_d)$; boundary value changed by a Δ.
6: $S_t m \leftarrow S_t \cup x_m$; modify the corresponding state feature variables with the new x_m.
7: $\vec{X_c} \leftarrow \vec{X_c} \cup \widehat{\mathbb{T}} \cdot predict(S_t m)$; get the counterfactual decision nodes by getting the counterfactual action and then traversing the tree.
8: **end for**
9: **return** $\vec{X_c}$

Just decision-tree policy:	Because Ally unit number (A_n) is 4 and Barracks number (B) is 1.
With action influence models:	Because ally unit number (A_n) is 4 and the goal is to have more Destroyed Units (D_u) and Destroyed buildings (D_b).

4.5.3 Contrastive Explanations from Counterfactuals

Counterfactuals explain events that did not happen—but could have under different circumstances. Counterfactuals are used to describe events from a 'possible world' and to contrast them with what happened in actuality. Embedding these counterfactuals in explanations can make the explanation more meaningful [36]. Naturally, an explanation given to a 'why not' question should compare the counterfactuals with the actual facts to form a *contrastive explanation* [156, 155]. For this reason, we concern ourselves with generating contrastive explanations from decision nodes and causal models.

We generate the counterfactual decision nodes using Algorithm 2, in which we find the decision nodes of the counterfactual action b by changing the decision boundary of the actual action b in the decision tree. We can now define *minimally complete contrastive* explanations for 'why not' questions using these counterfactual decision nodes.

Definition 4.5.2. Given the set of decision nodes $\vec{X_d} = \vec{x_d}$ for the action a from a decision tree $\widehat{\mathbb{T}}$, a *minimally complete contrastive* explanation for a *why not* question is a pair $\left(\vec{R} = \vec{r}, \vec{X_{con}} = \vec{x_{con}}\right)$, in which $\vec{R}$ is same as in Definition 4.5.1; $\vec{X_{con}}$ is such that $\vec{X_{con}}$ is the

maximal set of variables in which $\overrightarrow{X_{con}} = (\vec{X}_b = \vec{x}_b) \cap (\vec{X}_c = \vec{x}_c)$, where $\vec{X}_b$ gives the set of intermediate nodes of the causal chain of the counterfactual action b, and $\vec{X}_c$ is generated using the Algorithm 2. Values $\vec{r}$, $\vec{x}_c$ are contrasted using the actual instantiation $M_{\vec{V} \leftarrow \vec{S}}$ and counterfactual instantiation $M_{\vec{Z} \leftarrow \vec{S}_Z}$ from Definition 4.4.2. ■

Instead of just having the intermediate nodes of the causal chain of the *actual* action (as in Definition 4.5.1), we now get the set of intermediate nodes for for the *counterfactual* action from its causal chain. Then the intermediate nodes of the counterfactual chain is compared with the set of nodes we get from the Algorithm 2, to get the common set of nodes, of which the variable values will finally be contrasted.

As before, we explain Definition 4.5.2 using the adversarial StarCraft II task. Consider the question 'Why not action A_b', when the actual action is A_s, for which the explanation is generated as follows. We first get the decision nodes A_n and B having $<= 5$ and > 2 as the decision boundaries respectively. Then each decision boundary value starting with the node closest to the leaf node, is moved by a small Δ amount 0.01 and applied as the new feature value in the current state of the agent (B feature value will change to 1.99). We use this new state to predict the counterfactual action as A_b from the decision tree, and to get the counterfactual decision nodes (which remains the same). Next, we get the intersection of nodes in the causal chain of the counterfactual action A_b ($B \rightarrow A_n \rightarrow [D_u, D_b]$) with $\vec{X}_c$, which gives B as $\overrightarrow{X_{con}}$ with the actual value 3 and counterfactual value 1.99. Finally, these values are contrasted and appended with the reward nodes of the causal chain of A_b to generate the explanation. A graphical interpretation of this explanation is shown in Figure 4.5.

4.5.4 Learning Opportunity Chains

Explaining the behaviour of the agent using only the policy (or a surrogate policy) alone, even if the explanation is causal, has shortcomings as this does not consider that some actions might be chosen because they enable other actions. In this section we discuss how information on enabling actions can be used to form a more complete explanation.

In the context of reinforcement learning, we define a 'distal action' as the action that depends the most on the execution of the *current* action of the agent. The agent might not be able to execute the distal action unless some other action was executed first (i.e. *some* actions 'enable'

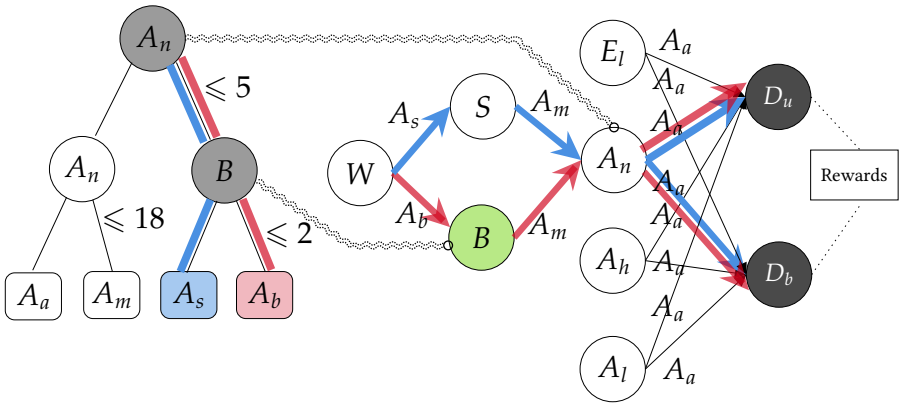


Figure 4.5: Visual example of the explanation for *Why not* A_b. Actual action and the actual causal chain is shown in blue, counterfactual chain and nodes are shown in red and the contrast node is shown in green.

the execution of other actions). For example, in the StarCraft II domain, the action 'train marines' cannot be executed until 'build barracks' action is executed. While it is possible to extract distal actions from environment dynamics and pre-conditions in a model-based system, for model-free RL agents, this remains a challenge. However, for the purpose of explanation, it is possible to provide an approximation and predict the distal action.

We use a many-to-one recurrent neural network (RNN) [202] as our prediction model $\widehat{\mathbb{L}}$ to approximate the distal action given a sequence of previous states and actions of the agent. We implement $\widehat{\mathbb{L}}$ with a fully connected hidden layer of 10 units, and a batch size of 100. For training data, we use the batch replay dataset D_t discussed in Section 4.2. We define a sequence as a state-action trace that ends in an action that is one of the last actions in a causal chain (e.g. in Figure 2, the last action of all causal chains in the 'attack' action). The output of the model $\widehat{\mathbb{L}}$ will be the distal action and its expected cumulative reward. Formally, this action prediction model can be written as: $\hat{y}_{t_{N+1}} = f(x_{t_1}, x_{t_2}, ..., x_{t_N}; t_1, t_2, ..., t_N)$, where x_{t_N} is the state-action pair (including state features) of the last action of a causal chain, $\hat{y}_{t_{N+1}}$ gives the distal action and the reward. Here, we use the immediate next action that lies in a particular causal chain as the ground truth. Note that even though we used an RNN to implement the prediction model, it is entirely possible to use other models to approximate the distal action. With the distal action prediction model $\widehat{\mathbb{L}}$ in hand, we now define *minimally complete distal* explanations for 'why' and 'why not' questions that incorporate causal nature to the explanations.

Definition 4.5.3. Given a *minimally complete contrastive* explanation, current action a and a prediction model $\widehat{\mathbb{L}}$, a *minimally complete distal* explanation is a tuple $\left(\vec{R} = \vec{r}, \vec{X_{con}} = \vec{x_{con}}, a_d\right)$, in which $\vec{R}$ and $\vec{X_{con}}$ do not change from Definition 4.5.2; and a_d gives the distal action predicted through $\widehat{\mathbb{L}}$ such that $a_d \in A \cap A_c$, where A is the action set of the agent and A_c gives the action set of the causal chain of current action a. ■

Informally, this simply prepends the predicted distal action to a minimally complete contrastive explanation generated through Definition 4.5.2 if the distal action exists in the causal chain of the current action. Consider the example 'Why not action *build_barracks* (A_b) , when the actual action is *train_marine* (A_m). This would yield the counterfactual decision node A_n (ally unit number) with the actual value 10 and the counterfactual value 5. When the predicted distal action is *attack* (A_a), we can generate the below explanation text using a simple natural language template. The causal explanation is generated with Definition 4.4.3 while the distal explanation is generated through Definition 4.5.3.

Causal Explanation: Because it is more desirable to do the action train marine (A_m) to have more ally units (A_n) as the goal is to have more Destroyed Units (D_u) and Destroyed buildings (D_b).

Distal Explanation: Because ally unit number (A_n) is less than the optimal number 18, it is more desirable do the action train marine (A_m) to *enable the action* attack (A_a) as the goal is to have more Destroyed Units (D_u) and Destroyed buildings (D_b).

Note that the Definition 4.5.3 can also be used in conjunction with the Definition 4.5.1 to generate distal explanations for 'why' questions.

4.5.5 Computational Evaluation

We use five OpenAI benchmarks [31] and the adversarial StarCraft II scenario (discussed in Section 5.1) to evaluate the task prediction [110] accuracy of our distal explanation model and compare against action influence models as a baseline. Task prediction can be used to predict what the agent will do in the next instance, and measures how faithful the surrogate policy is against the underlying policy.

Env - RL	Size	SE - Accuracy (%)			DP - Accuracy (%)	
		LR	DT	MLP	DP	DP_n
Cartpole-PG	4/2	83.8	81.6	86.0	96.83	97.10
MountainCar-DQN	3/3	69.7	57.8	69.6	88.66	86.75
Taxi-SARSA	4/6	68.2	74.2	67.9	82.44	86.19
LunarLander-DDQN	8/4	68.4	63.7	72.1	72.82	72.91
BipedalWalker-PPO	14/4	56.9	56.4	56.7	67.99	69.28
StarCraft-A3C	9/4	94.7	91.8	91.4	97.36	86.04

Table 4.3: Distal explanation model evaluation in 6 benchmark reinforcement learning domains that use different RL algorithms, measuring mean task prediction accuracy in 100 episodes after training. SE-structural equations (trained with LR-linear regression, DT-decision trees, MLP-multi layer perceptrons), DP-decision policy tree and DP_n-unconstrained decision policy tree.

We choose the benchmarks to have a mix of complexity levels and causal graph sizes (given by the number of actions and state variables). We train the RL agents using different types of model-free RL algorithms (see Table 4.3), using a high performance computer cluster node with 2 Nvidia V100 GPUs, 56GB of memory and 20 core CPU with 2.2GHz speed. All agents were trained until the reward threshold (to consider as 'solved') of the environment specification is reached.

We evaluate two versions of the distal explanation model, where one is a based on a depth limited decision tree with the number of actions (DP in table 4.3), other trained until all leaves are pure nodes (DP_n). Results summarised in Table 4.3 show our model outperforms task prediction of action influence models (with their structural equations trained by either linear regression (LR), decision trees (DT) or multi layer perceptrons (MLP)) in every benchmark, some by a substantial margin.

The benefit gained through unconstrained decision trees (DP_n) does not translate well into an increase in task prediction accuracy. We conclude that for the purpose of using distal models for explanation, a depth limited tree (DP) provide an adequate level of accuracy. Moreover, as a depth limited tree is likely to be more interpretable to a human, it is more suited for *explainability* and *explanation*.

Figure 4.6: StarCraft II Collaborative task scenario: The agent is controlling the leftmost section and the participant controls the right section (divided by the fissure)

4.6 Evaluation: Human Study

We consider human subject experiments to be an integral part of XAI model evaluation and as such conduct a human study with 90 participants. We consider two hypotheses for our empirical evaluation; 1) Distal explanation models leads to a improved *understanding* of the agent; and 2) Distal explanation models provide *subjectively* 'better' explanations. Our experiment involves RL agents that complete objectives in three distinct scenarios, which are based on the StarCraft II [235] learning environment. We first discuss these scenarios below.

4.6.1 Scenarios

In addition to the default scenario of the StarCraft II, we developed two additional scenarios as custom maps using the StarCraft II platform as a framework, that are better suited for explainability. Custom maps were made to add a more strategic nature to scenarios and in some cases to elicit cooperation from the interacting human. Note that these scenarios are completely different from the StarCraft II *game*. We only use StarCraft II assets as a simulation framework, similar to how e.g. a grid-world framework can be used to make many different scenarios. We also

release these maps with state and action specifications as test-beds for explainability research.

Adversarial In this scenario, the agent's objective is to build its base by gathering resources and destroy the enemy's base. The agent can build offensive units (marines) to attack the enemy's base and to defend its own base. This is the default objective in a normal StarCraft II game, but here we only use 4 actions for the purpose of the experiment rewards are given for the number of enemies and buildings destroyed (shown in Figure 4.4 b) as an action influence graph). During the experiment, the trained RL agent will provide explanations to the participant and the strength of the explanations are evaluated through task prediction.

Rescue This scenario is a custom map, where the agent's objective is to find a missing unit and bring it back to the base using an aerial vehicle. The agent also has to avoid or destroy enemy units during the rescue and aid the aerial vehicle using an armed unit. The agent has access to 5 actions, the reward is given for the number of missing units saved. The evaluation is done through task prediction as before.

Collaborative Task The collaborative task is fundamentally different from the previous scenarios, in that the participant has to help the agent to complete the objective. We made this task as a custom map (depicted in Figure 4.6) where the map is partitioned as the agent and human 'area'. The agent can perform 5 actions in this task, while the human can choose 4 actions to execute. The objective of the task is to build a series of structures that finally leads to the creation of an 'elite' unit, which the human has to transport to a base. The success of the task depends on the participant choosing to execute the action that best support the agent.

4.6.2 Experiment Design and Methodology

To investigate the two main hypotheses, we use a mixed design [121] (within subject and between subject) for our experiment. Every participant will be evaluated on the 3 independent variables which are 1) 'no explanations', where only a visual description of the agent behaviour is provided; 2) causal explanations generated with action influence models and 3) our distal explanation model. At a glance, the experiment has 3 phases where participants receive explanations from RL agents,

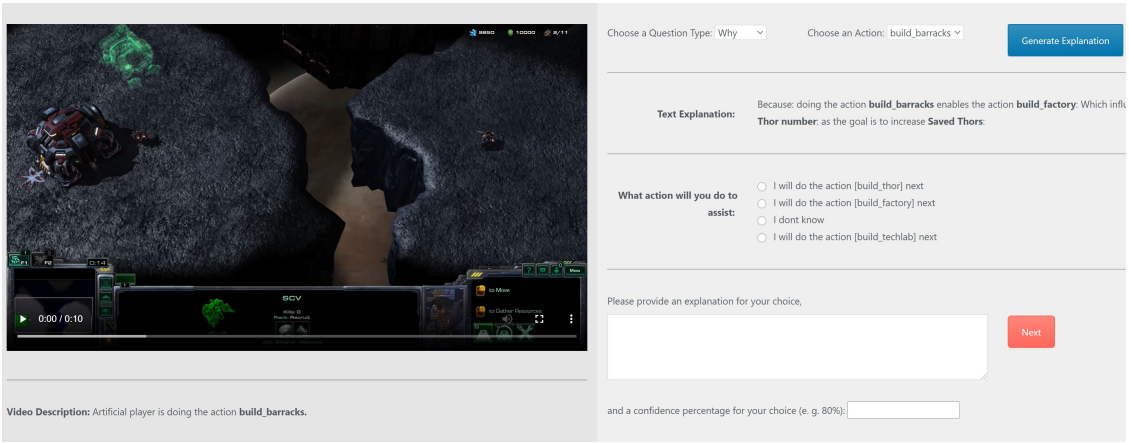


Figure 4.7: The web-based interface of the experiment showing the Collabrative Task.

subjectively evaluate the explanation and are then evaluated through task prediction [110] to gauge their understanding of the agent.

Task prediction is an effective measure that can peek into the mental model of an explainee to evaluate how successful the given explanation was in transferring the knowledge from the explainer [110, 156]. In task prediction, the participant is asked the question 'What will the agent do next?'. We use task prediction to evaluate the hypothesis 1) for the Adversarial and Rescue scenarios, and invert the question as to ask 'What would you do next?' in the Collaborative task. We investigate hypothesis 1) by employing the 5-point Likert *explanation satisfaction scale* of Hoffman et al. [110, p.39]. Explanation satisfaction is evaluated after each explanation and also at the end of the experiment which compares explanations of causal and distal models.

Experiment Design: We use *Amazon Mechanical Turk* (Mturk)—a crowd sourcing platform well known for obtaining human-subject data [35]—to conduct the experiments. A web-based interactive interface is used as the medium of interaction.

We first display the ethics approval obtained through a university, and after the participants' consent gather demographic information. We then show video clips of the agents solving the 3 StarCraft II scenarios that capture the behaviour of the agents. Each scenario has 4 distinct behaviours of the respective agent (around 10 seconds per clip). Every participant sees all three scenarios, and all three explanation types, but between participants, the combination of scenario and explanation type are mixed. For example, a participant may experience: Adversarial with no

explanations, Rescue with casual explanations and Collaborative with distal explanations. The order of these is randomised to control for ordering effects.

The first stage of the experiment involves training the participants to identify agents' actions using video clips of the agents performing those actions before the start of each scenario. In the Collaborative scenario, participants are trained to identify the actions they can use instead. After validating that participants can distinguish different actions through a question, the scenario will be presented.

The second stage lets the participants ask explanatory questions (in the form of why/why not *action*), after watching the agent's behaviour through the video clip. Participants can ask any number of questions and we did not control for a minimum number of questions; however, we incentivised participants to ask questions because they knew they would receive bonus payments for getting predictions correct later in the experimence. After each explanation video, participants are presented with the explanation satisfaction survey. For each explanation/scenario pair, each participant engages in 4 tasks.

The third stage involves evaluating the participants' 'understanding' of the agent through task prediction. Participants are presented with 4 new videos with different situations, and are asked what action the agent will do next, and can select one of the 4 options (which are 3 actions of the agent plus the option of 'I don't know'). Each participants makes predictions for 4 tasks. After this stage participant will move to the next scenario with a different explanation model and repeat from Stage 1 to 3. This is done until all the scenarios are encountered by the participant.

In the final stage, the participant is presented with 3 additional explanation videos (of the scenario they did for the no explanation condition), and is presented with causal explanations from action influence models and our distal explanation model *side by side*. We use Hoffman et al. [110, p.39]'s explanation satisfaction scale but this time as a movable slider that subjectively compares the two explanation.

Experimental Conditions: We ran the experiment with the above mentioned 3 independent variables (the explanation models), which resulted in 3 combinations of explanation model and scenarios, with participants seeing all 3 scenarios and all 3 explanation types. Each combination had 30 participants for a total of 90 participants in the experiment. Each participant is scored on the total number of correct task predictions out of 12 (4 each for each model-scenario

combination).

Each experiment ran approximately 50 minutes, and we compensated each participant with 8.5USD (a bonus compensation of 0.5USD was also given to participants for each point above 10). Participants were aged between 23 to 60 ($\mu = 38.1$), and of the 90 participants, 51 were male while 38 were female and 1 who did not provide an answer. Participants reported an average self-rated gaming experience and StarCraft II experience of 2.47 and 1.47 out of 5 (5-point Likert) respectively.

To ensure the quality of data from participants, we recruited only 'master class' workers with 95% or more approval rate. We controlled for language by only recruiting workers from the United States. We excluded the noisy data of users in 3 ways. First, we tested participants to ensure they had learnt about the scenario by asking them to identify actions shown in several videos. If the participant failed this, the experiment did not proceed (participants were paid a $2USD base amount). Second, we tracked how much time each participant spent viewing explanations and answering tasks. If this was regularly below a threshold of a few seconds, we omitted that participant from our results. Third, participants were required to explain their task predictions. If this text was gibberish or a 1-2 word response, we omitted that participant from the results. We filtered out 16 participants according to the above constraints to yield the final participant number of 90.

4.6.3 Results

We first discuss the results on hypothesis 1), where we investigate whether distal explanation models lead to a better understanding of the agent. We present the null hypotheses as $H_0 : P_N = P_C = P_D$ and the alternate hypothesis as $H_1 : P_D > P_N$ and $H_2 : P_D > P_C$, in which N, C, D corresponds to 'no explanation', causal and distal explanation models. Here, P denotes the proportions of the observed values of correct answers in task predictions by the participants.

We perform Pearson's Chi-squared test for the three StarCraft II scenarios and obtain the following values: Adversarial (p-value = 0.011, $X^2 = 13.00$), Rescue (p-value = 0.034, $X^2 = 10.40$) and Collaborative (p-value = <0.001, $X^2 = 35.47$). As the Chi-squared test was significant at the 0.05 level across the three scenarios, we investigate the pairwise differences between models using a z-test. We summarise the results in Table 4.4. From Table 4.4, considering the proportions

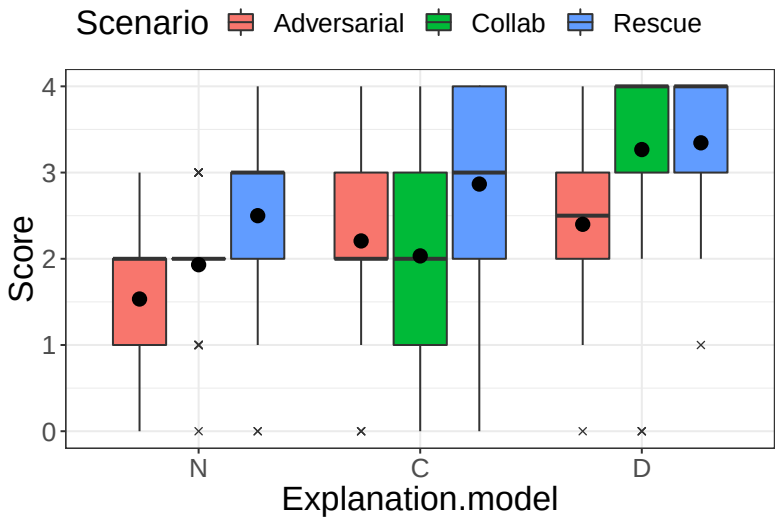


Figure 4.8: Box plot of task prediction scores of the explanation models across the StarCraft II scenarios (means are represented by bold dots)

(P) between model pairs, we can see that apart from Adversarial and Rescue scenarios for the D - C model pair, distal explanation models have statistically significant results at the 0.05 level between other combinations. Thus we accept H_1 for every StarCraft II scenario and accept H_2 only for the Collaborative scenario. We further test the validity of our results by employing a pairwise t-test which produce similar conclusions (results shown in Table 4.3. We illustrate these results as a box-plot in Figure 4.8. Clearly, the Collaborative scenario poses a much higher challenge to the participants, and results indicate that distal explanations perform better than other models in this task.

Explanation Quality: The second main hypothesis 2), evaluate whether distal explanations can provide *subjectively* better explanations. The corresponding null hypothesis is $H_0 : P_N = P_C = P_D$ and the alternative hypothesis is $H_1 : P_D > P_C$. P in this case P becomes the proportion of the observed values of the Likert scale data (using the survey of Hoffman et al. (2018, p.39)), where participants have rated as '5'. We consider four explanation quality metrics; 'Complete', 'Sufficient', 'Satisfying' and 'Understanding'. As before, we employ Pearson's Chi-squared test to see the significance of the above 4 metrics in the 3 StarCraft II scenarios, and obtain p-values < 0.01 for every condition. As Likert data are ordinal type data and the distribution of the data cannot be assumed to follow a normal distribution, we use a non-parametric test, Chi-squared test along

Table 4.4: Pairwise differences with a z-test for proportions for each model-pair and pairwise t-tests in the three StarCraft II scenarios in task prediction scores, considering the correct response.

Model (m_1-m_2)		Z-test			T-test	
	Scenario	X^2	p-value	Prop (m_1|$m2$)	p-value	t-stat
C - N	Adversarial	5.437	0.019	0.53 | 0.38	0.056	-1.988
	Rescue	2.283	0.130	0.71 | 0.62	0.169	-1.408
	Collaborative	0.416	0.518	0.50 | 0.46	0.537	-0.623
D - N	Adversarial	11.269	**<0.001**	**0.60 | 0.38**	**<0.001**	**-3.791**
	Rescue	9.931	**0.001**	**0.80 | 0.62**	**0.010**	**-2.750**
	Collaborative	31.966	**<0.001**	**0.81 | 0.46**	**<0.001**	**-4.761**
D - C	Adversarial	1.085	0.297	0.60 | 0.50	0.325	-1.000
	Rescue	2.784	0.095	0.80 | 0.71	0.221	-1.249
	Collaborative	25.511	**<0.001**	**0.81 | 0.50**	**<0.001**	**-4.367**

Table 4.5: Pairwise differences with a z-test for *explanation quality* metrics in models Distal (D) vs Causal (C), data where participants rated '5'.

Metric	Scenario	X^2	p-value	Proportions (D|C)
Complete	Adversarial	3.267	0.070	0.56 | 0.45
	Rescue	0.074	0.785	0.33 | 0.35
	Collaborative	11.428	**<0.001**	**0.40 | 0.20**
Sufficient	Adversarial	6.020	**0.014**	**0.56 | 0.40**
	Rescue	0.018	0.892	0.35 | 0.34
	Collaborative	15.55	**<0.001**	**0.41 | 0.18**
Satisfying	Adversarial	1.085	0.297	0.46 | 0.40
	Rescue	1.528	0.216	0.29 | 0.36
	Collaborative	9.981	**0.001**	**0.42 | 0.23**
Understanding	Adversarial	1.377	0.240	0.46 | 0.39
	Rescue	0.071	0.788	0.35 | 0.37
	Collaborative	15.31	**<0.001**	**0.42 | 0.19**

with proportion analysis [216]. As there are significant differences between explanation models on explanation quality, we reject H_0 and conduct a pairwise z-test. We summarise the results in Table 4.5. Figure 4.9 captures the Likert scale data distribution across models and scenarios. Though it is visually evident that distal explanation quality across the compared metrics has a positive trend, from the three scenarios, only the Collaborative task yields significant results for every explanation quality metric (see Table 4.5). Thus we accept H_1 only for the Collaborative scenario.

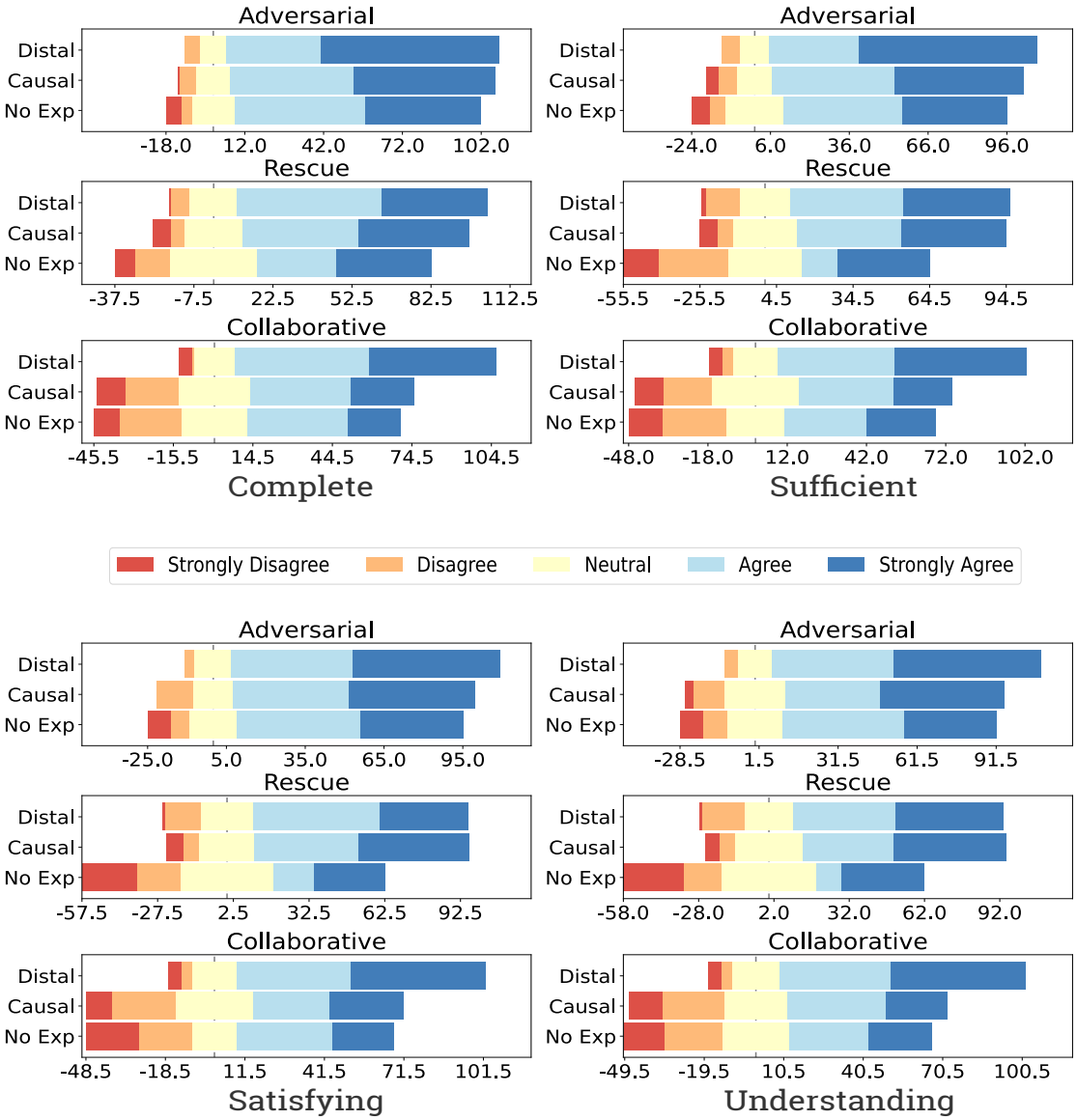


Figure 4.9: Likert scale counts of explanation quality metrics and how they vary across explanation models and scenarios. X-axis represent the total counts each Likert category received, adjusted to represent 0 as the midpoint.

Discussion: The results we obtained for *explanation quality* mirror the results in task prediction. Intuitively this makes sense as participants are more inclined to rate an explanation 'good' if they feel they have a better 'understanding' of the agent. Further investigations are needed to explore why distal explanations perform substantially better in human-agent collaborative tasks.

To investigate whether the knowledge of the StarCraft II game had any impact on task prediction scores, we perform a Pearson's correlation test between task prediction and StarCraft II experience (self-report in a 5-point Likert scale). The obtained values (t = 1.515, p-value = 0.133) indicate that there is no statistically significant correlation between scores and StarCraft II experience. Although our experiment was based on the StarCraft II environment, we used custom maps and scenarios that are different from the game. Thus the results of the correlation test is plausible.

One weakness of our model is the need for a causal graph that is faithful to the problem, in order to learn the *opportunity chains*. For the purpose of this work, we hand-crafted the causal graphs for StarCraft II scenarios and the 5 RL benchmarks. While our hand-crafted models can be verified easily with data, we acknowledge that it may become infeasible in larger domains. We view generating a causal graph a distinct problem than generating explanations *using* a causal graph. As such we propose this as our immediate future work.

Limitations of the experiment design: Although we used scenarios that have different objectives than the standard Starcraft II game, familiarity with the game's concepts may have had some impact on the scores even if it is not significant. Participants also may have had commonsense knowledge about such scenarios (in particular rescue and adversarial) that can affect their judgments. Our results should be generalisable across similar scenarios in different *domains*, though further experiments are needed to evaluate the generalisability across different scenarios (e.g. path planning, manufacturing).

4.7 Conclusion

We introduce a *distal explanation* model for model-free reinforcement learning agents that can generate explanations for 'why' and 'why not' questions. These models learn *opportunity chains* (in the form of A enables B and B causes C), and approximate a future action that *enables* due to the current action of the agent. Our motivation comes from insights gained through a human-agent experiment, in which we analysed 240 human explanations. Participants in this study frequently referred to future action that depend on the current action of the agent, which conform to the definition of opportunity chains. To learn opportunity chains at the training phase in reinforcement learning we make use of *action influence models* to extract causal chains

and represent the approximated policy of the agent in a decision tree policy. In contrast to action influence models that use structural causal equations to generate *contrastive* explanations, we use the decision policy in conjunction with causal chains to improve the accuracy of *task prediction*. We evaluate our approach in 6 RL benchmarks on task prediction. We then undertake a human study with 90 participants to investigate how the distal explanation model perform in task prediction and *explanation quality* metrics in three custom scenarios built using the StarCraft II platform.

While results indicate a significantly better performance of distal explanations compared with two other explanation models in collaborative situations, further research is needed to understand the impact this technique may have on other types of scenarios. One weakness of our model is the need of knowing the causal structure of the domain beforehand. Though this can be mitigated by using existing causal discovery methods, the reinforcement learning setting provides a unique opportunity to learn causal graphs that are better suited for explanation through influencing the exploration of the agent. Chapter 5 of this thesis discuss discovering the action influence structure at the training time of the agent.

Chapter 5

Action Influence Discovery for Explainable Reinforcement Learning

5.1 Introduction

In this chapter, we introduce a novel way to learn an explainable model from the RL agent interaction data of the environment, that is based on action influence and causal relationships of the environment variables. In Chapters 3 and 4, we focused on how to generate explanations when the agent's action influence structure is known. Here we develop methods that can learn this influence structure end-to-end only through the agent's previous state-action traces.

When explaining an RL agent, two distinct approaches can be taken. First, the agent's action can be explained (commonly known as local explanations). Second, the agent's policy can be explained, referred to as global explanations (e.g. summarising agent strategies [10]). Local explanations help the explainee to understand the reasoning of a particular action while global explanations make the overall behaviour of the agent more intelligible. A major shortcoming of both local and global explanation generation strategies for RL agents is the need of requiring an underlying *explainable* model, often handcrafted by domain experts. Here we introduce a mechanism that can learn an explainable model autonomously through the agent's interactions with the environment, that can be used to generate local explanations.

The notion of action influence has been used to understand agents' goals that can mitigate safety concerns [71, 70], and improve coordination in multi-agent settings [244]. Influence diagrams have also been used to model the agent's beliefs [217] and mental models [78]. *Action influence models* were introduced in Chapter 3, which captures how can affect the endogenous variables

(variables that are inside the agent's model) and their causal relationships. In Chapters 3 and 4 we discussed how Action influence models can be successfully used to generate explanations, and have been shown to improve the human's understanding of the agent. One of the main shortcomings these models have is the requirement of a hand-crafted action influence structure, which capture the causal relationships between agent's state variables and how the agent's actions influence them. As environments become larger and more complex, hand-crafted models can become infeasible and erroneous. To this end, in this Chapter, we introduce an architecture that *learns* the action influence structure, only using the RL interaction data with the environment (i.e. state-action traces of the agent).

The action influence learning architecture is composed of 3 stages; first, inferring the influences of actions on state variables and encoding them; second, learning the causal relationships between these encoded algorithms using causal discovery; and third, decoding the learned graph structure into an action influence graph. We use the causal discovery method of Zhu, Ng, and Chen [257] and adapt it to our architecture to learn the causal relationships of RL agents' state variables. Note that this causal discovery algorithm can plausibly be substituted with another state-of-the-art method, given it can be adapted to handle RL agent data streams. More importantly, though we only focus and discuss the method in the context of model-free RL agents, our architecture can be used with many other sequential decision making agents that have state-action based modelling (e.g. planning agents).

We evaluate the action influence discovery algorithm on 5 RL benchmark domains. Here, open AI domains and 2 different StarCraft II domains were selected, considering the number of state features, actions, and the number of connections (action influences). We measure the correct number of edges that the algorithm discovered. Here, a *correct* edge has to match the causal relationship between variables (presence of a causal relationship and its direction) *and* the correct action that influences the variables, against a ground truth action influence model. In addition, we also report on missing, extra edges and incorrectly labelled actions. Results indicate that our action influence discovery architecture can learn the action influence structure reasonably well to use them in generating explanations.

The main contribution of this Chapter will be the action influence discovery architecture, with its methods and algorithm. This Chapter answers the **RQ3** of the thesis, and forms an important

base for the explanation generation methods described in Chapters 3 and 4.

5.2 Background

Action influence models are causal models augmented with the actions of an agent. We briefly discuss action influence models in the context of an RL agent below.

5.2.1 Preliminaries

Action influence models are defined for RL agents based on Markov Decision Processes (MDP), and we make use of the standard definition. Action influence models describe the *causal* relationships between state variables of the agent that are *influenced* by the agent's actions. This model can be represented using a directed acyclic graph (DAG) with state variables as the nodes and edges annotated with the agent's actions. In contrast, vanilla causal models does not have labeled edges because the influence between variables is not an agent action.

Informally, action influence models gives a model that capture the *causal structure* and the *causal effects* that might exist between the state variables of the agent. Graphically, the causal structure between two or more variables when influenced by an action is depicted by a directed labeled edge in a DAG. The causal effect between two or more variables is captured by structural causal equations [92]. Consider the following example.

Example 5.2.1. Consider the *coffee task* [29] where a robot has to deliver coffee to a user. The state consists of six binary variables, robot location (L), robot is wet (W), robot has umbrella (Umb), raining (Rn), robot has coffee (C) and user has coffee (Usr). Actions of the robot are *go*, *buy coffee*, *get umbrella* and *deliver coffee*. We can formulate an action influence model M by identifying the set of functions $\mathcal{F}$ that describe causal relationships of state variables. Assuming there is only one such function, we can define it as $F_{Usr.deliver-coffee} = C + L$. This implies that the variable 'user has coffee' is causally influenced by variables 'robot has coffee' and 'robot location' when the action *deliver coffee* is executed by the robot. Prior knowledge of the causal structure is needed to formulate this function (Here, Usr is influenced by L and C through the action *deliver coffee* has to be known beforehand). Causal effect can be approximated by training these functions as regressors.

In Chapter 3 we discussed how the causal effect of action influence models can be approximated in model-free RL agents. While this allowed action influence models to generate causal explanations, one notable weakness is the reliance in prior knowledge of the causal structure. In the work presented in Chapter 3, the causal influence structure needs to be handcrafted using domain knowledge. Importantly, the work described in Chapter 3 only learns the structural causal *equations* for a given action influence structure, whereas in this work we focus on discovering the innate action influence structure.

5.2.2 Generating Explanations

We use the explanation generation method proposed in Chapter 3 that can generate 'why' and 'why not' explanations using an action influence graph. Note that this paper only focused on discovering the action influence graph, and not on approximation of the structural equation. Further, other types of explanation generation techniques can also be applied using the causal structure (e.g. Distal explanation generation discussed in Chapter 4).

5.3 Action Influence Discovery

In this section, we introduce a causal influence discovery model that learns the structure of the causal relationships that are influenced by the agents actions, without any prior knowledge about the influence structure of the domain.

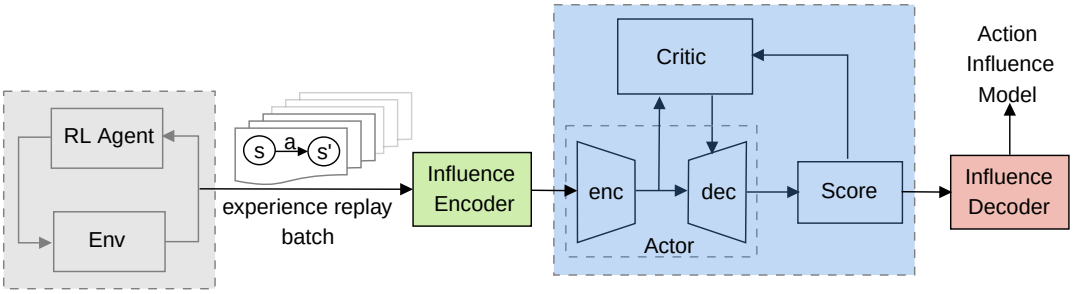


Figure 5.1: An overview of the Causal Influence Discovery Model

The action influence discovery problem can be decoupled into three main components. Figure 5.1 shows an overview of this model. We define the architecture using a 4-tuple $\mathcal{M} = (D, E, C, D)$.

The agent's interactions with an environment (shown in grey) is used to produce experience replay [136] batches, given by D. E encodes the action influences into variables. In contrast to traditional causal discovery methods that rely only on observational data (i.e. variable changes), we encode the influence an action can have on the variable change (shown in green). C denotes the causal discovery algorithm. We use Zhu, Ng, and Chen [257]'s reinforcement learning based causal discovery method (shown in blue) to find the causal graph structure $\mathcal{G}_c$ that best describes the encoded influence data. Importantly, our model can use an alternate general causal discovery method as the influence encoding is decoupled from causality learning. The learned best causal structure would then be decoded into an action influence graph $\mathcal{G}_a$ (shown in red), given by D. Note again that we can substitute the causal discovery algorithm C if better algorithms are derived in the future literature. For the propose of the empirical evaluation, Zhu, Ng, and Chen [257]'s causal discovery algorithm is adapted to handle RL agent data. We discuss these components in detail below.

5.3.1 Influence Encoder

Given an experience replay dataset $D = \{s, a, s'\}_{t=m}^{t=n}$ of a RL agent, where m, n gives the time step range of the batch and $e = (s_t, a_t, s_{t+1})$ gives the current state, the action and the next state in an agent time step t, we are concerned with encoding the action influence effects. In Figure 5.2, the top half represents how the action influence is encoded between state-action transitions. This matrix like structure captures the action influence by generating a new dataset with influence encoded variables.

When the agent transition from s_t to s_{t+1} by executing the action a_t, some subset of state variables $\mathcal{I} = \{x_1, x_2 ... x_p\}$ of s_{t+1} will have altered values from the values in s_t (e.g. see darkened cells in Figure 5.2). We can then generate a new dataset $D_c = \{s_i\}$, where $s_i := f(x_i.a), i = 1, 2.., p \wedge x \in \mathcal{I}$. Dataset D_c can be fed into a causal discovery algorithm to find the causal relationships between variables. This will produce a causal graph similar to the one given in Figure 5.2. Note that in contrast to a causal graph one would discover through observational data, an action influence causal graph's nodes are associated with an action.

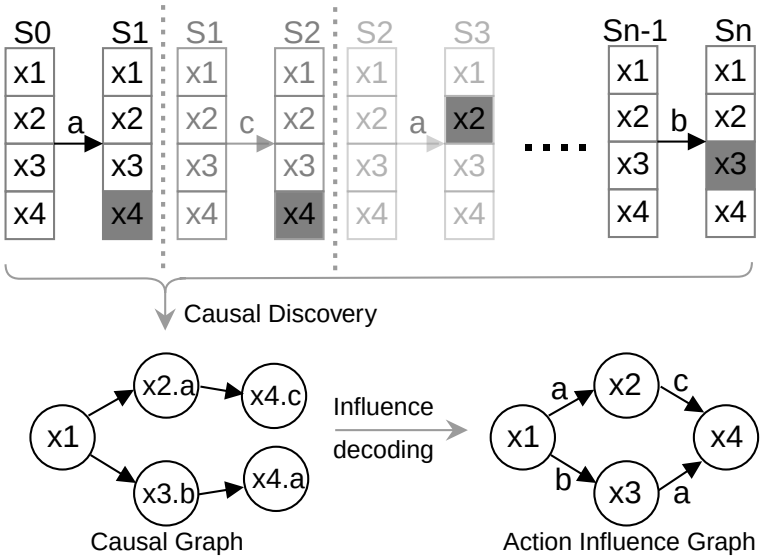


Figure 5.2: Encoding action influence (darkened cells indicate the state variable(s) influenced by the action) from RL experience replay batches for causal discovery and decoding to generate the Action influence graph.

5.3.2 Neural Architecture

We closely follow the neural architecture introduced by [257] and use the dataset D_c generated above to create the observational space. The purpose of the architecture is to generate a binary adjacency matrix $A \in \{0,1\}^{n \times n}$. Here, n denotes the action encoded feature nodes (e.g. $x2.a$ in Figure 5.2. The equivalent directed graph would be ensured for acyclicity and will be selected based on a scoring function descried below.

5.3.3 Encoding and decoding the graph

This section describes how the action influence dataset is translated into a graph-like structure. Note that this encoder-decoder model is distinct from the action influence decoder discussed above. The purpose of the encoder-decoder module is to generate a graph structure that represents the agent input data. Transformers [232] have been successful in representation learning in domains like natural language processing computer vision. We use the architecture proposed by Vaswani et al. as *neural encoder* that takes the dataset D_c described above as the input.

Two such encoders, enc_i and enc_j where $enc_{i,j} = 1,2,3...,n : dimension = d_{enc}$ is used as the input for the decoder, given by; $g_{ij}(W_1, W_2, u) = u^T tanh\left(W_1 enc_i + W_2 enc_j\right)$, where

$W_1, W_2 \in \mathbb{R}^{d_{enc} \times d_h}, u \in \mathbb{R}^{d_h}$ denotes the trainable parameters. We generate the adjacency matrix A each element at a step, as opposed to row-wise generation. Each element g_{ij} would be sampled using a probability distribution (Bernoulli) with probability $\sigma\left(_{ij}\right)$, where the probability indicates the likelihood of a directed edge from element x_i to x_j. Element (i, i) is masked to avoid cycles.

5.3.4 Searching for the Influence Graph

The encoder-decoder architecture described above would represent adjacency matrices where we need to select the one that best describes the action influence and causal relationships of the agent's model. Recent work of [257] found success in combining traditional score based methods with reinforcement learning to find the best graph. In this work, we apply the same search method to find the best influence graph.

5.3.5 Score Function

We use the well known Bayesian Information Criterion (BIC) [204] as the score function. BIC for a graph $\mathcal{G}$ is given as, $BIC\left(\mathcal{G}\right) = -2\log p\left(X; \hat{\theta}, \mathcal{G}\right) + d_\theta \log m$, where $\hat{\theta}$ gives the maximum likelihood estimator and d_θ gives the dimension of θ. Assuming linearity in the causal relationships of variables and assuming Gaussian additive noise we can rewrite the above as [257],

$$BIC\left(\mathcal{G}\right) = \sum_{i=1}^{d} \left(m \log\left(RSS_i/m\right)\right) + \left(no.edges\right) \log m$$

Here, RSS_i is the residual sum of squares for the i-th feature (with x_i^k denoting ith element in the kth observed sample and $\hat{x}_i^k$ giving the estimation of x), given by $\sum_{k=1}^{m}\left(x_i^k - \hat{x}_i^k\right)^2$.

5.3.6 Reinforcement Learning for Search

We employ the actor-critic architecture to formulate the search. This is represented visually in Figure 5.1 (in blue) in coupled with the encoder-decoder (as the actor). We describe the reward and the objective function of the agent below. The RL agent's reward constitutes of the score

function and additional penalty terms.

$$R = -\left[BIC\left(\mathcal{G}\right) + \lambda_1 I\left(\mathcal{G} \notin DAGs\right) + \lambda_2 h\left(A\right)\right]$$

Here, $\mathcal{G}$ gives the current graph that is considered and the $DAGs$ give the set of graphs that is scored. We use the same penalty terms (λ_1, λ_2) used by [257] to ensure the acyclicity of the selected graph. Penalty term is also used in conjunction with $h\left(A\right)$ introduced by [254], where $h\left(A\right) := trace\left(e^A\right) - d = 0$. Here, e^A gives the matrix exponential of A. I gives the indicator function and BIC gives the score function. We can now write the objective function of the agent as below,

$$J\left(\psi \mid s\right) = \mathbb{E}_{A\sim\pi(.|s)}\{R\}$$

Here, s is the state input given by the experience replay dataset D_c and is constructed by drawing random samples. We use policy gradient for the optimisation and use the REINFORCE [220] algorithm to obtain the gradient $\nabla\psi J\left(\psi \mid s\right)$. We use the Adam optimiser [124] to train the critic network using n samples from s as a batch. We follow the network architecture of [257] for the critic with a 2-layer feed forward network with ReLU units, with the input $\{enc_i\}_{i=1}^{d}$. Means squared error is used to for the error between the true rewards of the critic and the predictions.

Training this RL agent can be done at the run-time of the of explainable agent using the encoded action influence dataset. Output for this search would be a causal graph (a DAG) without the actions labeled (as seen in Figure 5.2), with action influence embedded within the nodes. We can now decode this graph to generate action influence graph.

5.3.7 Decoding the Action Influence Model

A node of the encoded graph $\mathcal{G}_e$ is denoted by $X_i \cdot a$, where X gives the feature with its associated action $a \in$ agent's action set. We follow a simple pruning method to label the edges and combine nodes of the graph. For every node $X_i \cdot a$ of the graph $\mathcal{G}_e$, every incoming edge e will be annotated with the action a. After the action annotation, if there exist node(s)X_i such that $i = j; 1, 2, ...p$, combine the nodes into one (while combining edges as well). As an example, from Figure 5.2, we can see the node $X_4 \cdot a$ and $X_4 \cdot c$ can be combined to the node X_4 while annotating the incoming edges from node X_2 as c and X_3 as a respectively. We get the action influence graph

$\mathcal{G}_a$ as the output of the pruning and annotating process. We can now use the action influence graph to generate explanations.

5.4 Empirical Evaluation

We carry out empirical experiments using 5 different RL domains and report on accuracy against a against ground truth models. We chose domains in the well known Open AI gym environments and custom made scenarios in the StarCraft II domain. Computational times for different domains are also compared.

5.4.1 Domains

Selection of domains were based on the complexity of the environment and the number of features (agent's endogenous variables) and the number of actions of the agent. We note that the domains were also chosen based on the ease of access to the ground truth action influence graph that can be inferred by the simulator.

OpenAI gym environments: From the openAI gym environments we used the Lunar-lander (features - 8, actions - 4), Taxi (features - 6, actions - 4) and the Cart-pole (features - 4, actions - 2). Ground truth action influence graphs we obtained through the consultation of the simulator code.

StarCraft II environments: StarCraft II is a real-time strategy game and a well-known RL playground for evaluating agent capabilities and is also used for explainability research. We used 2 different agents for the default StarCraft II map with one having 9 features, 4 actions and the other having 15 features, 8 actions. We obtained the ground truth through the simulator code and by consulting StarCraft II build trees.

In all of the domains, reply data was gathered in batches of 500 steps (i.e. batches of 500 interactions with the simulator), and our model is then trained interactively with incoming batches. In total, we gathered 500000 data points of state-action traces for each domain.

5.4.2 Measurements

To evaluate the generated action influence models against ground truth models, we introduce two measurements. These measures are based on causal discovery metrics but are augmented to handle actions.

Correct directed labeled edges: This metric count the correctly labelled directed edges of $\mathcal{G}_a$ against the ground truth action influence graph. An edge is correct iff the annotated action and the direction is correct.

Structural action hamming distance: This metric is based on structural hamming distance (SHD) [229]. Here we measure the difference between $\mathcal{G}_a$ and the ground truth action influence graph by counting the number of missing edges, extra edges and incorrectly labeled directed edges.

5.4.3 Results

Table 5.1 reports the results of our action influence discovery algorithm against the ground truth action influence models using the aforementioned metrics.

Domain	actions/features	correct vs no. g.truth edges	missing edges	extra edges	incorrect actions
Cart-pole	4/2	4/6	2	2	2
Taxi	4/5	10/16	6	2	0
Lunar Lander	8/4	22/40	12	10	4
StarCraft II v1	9/4	8/12	2	2	2
StarCraft II v2	15/8	12/20	4	4	2

Table 5.1: Structural action hamming distance (columns 4-6) and correctly inferred edges vs no. of ground truth edges edges of the generated action influence graphs.

From the table, all 5 domains yield comparable results to Zhu, Ng, and Chen [257]'s method that has a similar number of nodes (features). We note that direct comparison with the usual causal discovery benchmarks is not possible, due to the need for actions in action influence models. Further, to have a correct edge in action influence models, the action needs to match in addition to the edge and the edge direction (that represent the causal relationship). This is in contrast to the correct edge measurement of causal discovery methods, where only the direction and the presence of an edge is considered.

In the 5 domains used, considerable variations exist. While some domains have small action

and feature spaces (e.g. cart-pole), complex domains like StarCraft II has consist of large action and feature spaces. Importantly the number of actions and features (which makes up the nodes in the action influence graph) alone is a not good metric to measure the complexity of action influence models. The main objective of the action influence discovery algorithm is to correctly identify influencing action(s) in a causality related to two or more variables. As such the number of connections (with their actions) is the metric that we use here to compare against a ground truth model (Table 5.1 column 3). For example, from Table 5.1, in the Lunar Lander domain, we can see that the number of actions and features are not the highest of the tested. But this domain has the largest number of connections, due to the actions having influence over multiple feature variables at a connection. Over the 5 domains, the action influence discovery algorithm manages to correctly identify more than 50% of action influences. The performance of the algorithm is higher in domains like StarCraft II due to the clearly defined influence structure. Another important metric to consider is the number of incorrect actions. From the Table 5.1, our algorithm show a low number of incorrect actions. That is, in most cases, the algorithm manages to correctly identify the influencing action given the correct causal relationship and its direction. Though this is encouraging from the discovery of action influences, it also implies that the finding of the correct action influence graph can depend on the underlying causal discovery method used. We leave handling other types of causal discovery methods and incorporating them with action influence discovery to future work. Missing edges and extra edges given in Table 5.1 also largely depend on the causal relationships of the variables and the causal discovery method. Here an extra edge is an edge that does not exist in the ground truth graph. Though this edge will be labelled with some action, this does count towards the incorrect actions metric (i.e. an action is incorrect if it is labelled erroneously to different action to the action influence connection in the ground truth graph). Overall, the results indicate that the action influence algorithm can generation plausible action influence model structure from the RL agent data, which subsequently can be used to generate causal explanations using the work described in Chapters 3 and 4.

Accuracy of the structure of the action influence model can have an impact on the faithfulness of the generated explanations. Extra edges, missing edges and incorrect actions can all impact the accuracy of the explanation. These incorrect actions have the highest severity for explanation faithfulness. In Chapter 3, explanation generation methods do not make use of the full causal

chain, thus having several missing edges or extra edges would not impact the final explanation drastically. Note that missing edges can make some explanations impossible to generate (if the explainee ask a direct question about that edge). Incorrect actions can make explanations unfaithful as they can include incorrect causal chains. In our evaluation, we note that the incorrect actions discovered remain low compared to other metrics.

5.5 Conclusion

Causal explanations generated via Action Influence Models have been shown to increase the intelligibility of reinforcement learning agents. The main drawback of these models is the need for the action influence structure that exists between the environment and the agent, which capture the causal relationships between the agent's variables (i.e. state features) and how the agent's actions can influence those variables. Handcrafting this structure is time-consuming and is prone to human errors and might not be feasible in large domains. This chapter presents a novel action influence discovery algorithm that learns this structure using the replay data (state-action traces) of an RL agent, without interventions from a human expert.

In this chapter, we developed an architecture encode the action influences to learn the underlying causal structure. The agent's replay dataset is first encoded according to action influences, casting the influences as additional variables (in addition to features). Causal discovery is then performed using actor-critic reinforcement learning model to search for viable causal structure, using the BIC scoring method. The best causal graph is then decoded back to an action influence graph. Results indicate that the action influence structure discovered is accurate in how the actions are labelled, which can then be used to generate causal explanations.

Several future directions exist in improving the action influence structure for the purpose of causal explanation. Action influence structure can be learned at different levels of abstraction. In this work, we only considered the basic level of abstraction that exist within the data, though in generating explanations, the levels can differ between explainees. We expand upon this further in Chapter 7 Future work section. Another possible direction is using the action influence structure in the training process of the RL agent, where the agent's exploration can be guided by the information available in the action influence structure.

Chapter 6

An Interaction Protocol for Explainable Agents[1]

This chapter discuss how the social process of the explanation can be quantified into a grounded framework that is based on data from human explanation dialogues. We investigate the interaction between an explainer and an explainee and investigate the structural aspects of an interactive explanation to propose an interaction protocol, and follow a bottom-up approach to derive the model by analysing transcripts of different explanation dialogue types with 398 explanation dialogues. We use grounded theory to code and identify key components of an explanation dialogue. We formalise the model using the agent dialogue framework (ADF) as a new dialogue type and then evaluate it in a human-agent interaction study with 101 dialogues from 14 participants. Our results show that the proposed model can closely follow the explanation dialogues of human-agent conversations.

6.1 Introduction

In scenarios where people are required to make critical choices based on decisions from an artificial intelligence (AI) system, it is important for the system to able to generate understandable explanations that clearly justify its decisions. An appropriate explanation can promote trust in the system, allowing better human-AI cooperation [234]. Explanations also help people to reason about the extent to which, if at all, they should trust the provider of the explanation.

[1]This chapter is adapted from the published article: "A Grounded Interaction Protocol for Explainable Artificial Intelligence." Proceedings of the 18th International Conference on Autonomous Agents and Multi-Agent Systems. 2019.

As Miller [157, pg 10] notes, the process of Explanation involves two processes: (a) a *Cognitive process*, namely the process of determining an explanation for a given event, called the *explanandum*, in which the causes for the event are identified and a subset of these causes is selected as the explanation (or *explanans*); and (b) the *Social process* of transferring knowledge between explainer and explainee, generally an interaction between a group of people, in which the goal is that the explainee has enough information to understand the causes of the event.

However, much research and practice in explainable AI use the researchers' intuitions of what constitutes a 'good' explanation rather than basing the approach on a strong understanding of how people define, generate, select, evaluate, and present explanations [157, 158]. Most modern work on Explainable AI, such as in autonomous agents [246, 32, 41, 75] and interpretable machine learning [63], does not discuss the interaction and the social aspect of the explanations. The lack of a general interaction model of explanation that takes into account the end-user can be attributed as one of the shortcomings of existing explainable AI systems. Although there are existing conceptual explanation dialogue models that try to emulate the structure and sequence of a natural explanation [12, 238], we propose that further improvements will come from an empirically driven study of explanation.

Explanation naturally occurs as a continuous interaction, which gives the interacting party the ability to question and interrogate explanations. This allows the explainee to clear doubts about the given explanation by further interrogations and user-driven questions. Further, the explainee can express contrasting views about the explanation that can set the premise for an argumentation-based interaction. This type of iterative explanation can provide richer and satisfactory explanations as opposed to one-shot explanations. Note that we are not claiming that AI explanations are necessarily textual conversations. These interactions, questions, and answers can occur as part of other modalities, such as visualisations, but we believe that such interactions will follow the same model.

Understanding how humans engage in conversational explanation is a prerequisite to building an explanation model, as noted by Hilton [105]. De Graaf [59] note that humans attribute human traits, such as beliefs, desires, and intentions, to intelligent agents, and it is thus a small step to assume that people will seek to explain agent behaviour using human frameworks of explanation. We hypothesise that AI explanation models with designs that are influenced by

human explanation models have the potential to provide more intuitive explanations to humans and therefore be more likely to be understood and accepted. We suggest it is easier for the AI to emulate human explanations rather than expecting humans to adapt to a novel and unfamiliar explanation model. While there are mature existing models for explanation dialogs [238, 237], these are idealised conceptual models that are not grounded on or validated by data, and seem to lack iterative features like cyclic dialogues.

In this Chapter our goal is to introduce a dialogue model and an interaction protocol that is based on data obtained from different types of explanations in actual conversations. We derive our model by analysing 398 explanation dialogues using grounded theory [83] across six different dialogue types. Frequency, sequence and relationships between the basic components of an explanation dialogue were obtained and analyzed in the study to identify locutions, termination rules and combination rules. We formalize the explanation dialogue model using the *agent dialogue framework* (ADF) [150], then validate the model in a human-agent study with 101 explanation dialogues. We propose that by following a data-driven approach to formulate and validate, our model more accurately defines the structure and the sequence of an explanation dialogue and will support more natural interaction with human audiences than explanations from existing models. The main contribution of this chapter is a grounded interaction protocol derived from explanation dialogues, formalized as a new atomic dialogue type [240] in the ADF.

We first discuss related work regarding explanation in AI and explanation dialogue models, then we outline the methodology of the study and collection of data and its properties. We then present the analysis of the data, identifying key components of an explanation dialogue and gaining insight to the relationships of these components, formalising it using ADF and comparing with a similar conceptual model [25]. We then describe the human-agent study and present the validation of the model. We conclude by discussing the model with its contribution and significance in explainable AI.

6.2 Related Work

Explaining decisions of intelligent systems has been a topic of interest since the era of expert systems, e.g. [44, 120]. Early work focused particularly on the explanation's content, responsiveness and the human-computer interface through which the explanation was delivered. Kass and

Finin [120] and Moore and Paris [162] discussed the requirements a good explanation facility should have, including characteristics like "Naturalness", and pointed to the critical role of user models in explanation generation. Cawsey's [39] EDGE system also focused on user interaction and user knowledge. These were used to update the system through interaction. So, in early explainable AI, both the cognitive and social attributes associated with an agent's awareness of other actors, and capability to interaction with them, has been recognized as an essential feature of explanation research. However, limited progress has been made. Indeed recently, de Graaf and Malle [59] still find the need to emphasize the importance of understanding how humans respond to Autonomous Intelligent Systems (AIS). They further note how humans will expect a familiar way of communication from AIS systems when providing explanations.

Though there exist conceptual models of dialogical explanation [238, 239], these methods are not ground in human-agent explanations from a XAI context. The reader can examine section 2.6 for a more conclusive discussion on dialogical models and their limitations.

6.3 Methodology

To address the lack of a grounded explanation interaction protocol, we studied real conversational data of explanations. This study consists of data selection and gathering, data analysis, and model development, and then validation in a lab-based simulated human-agent experiment.

We designed a bottom-up study to develop an explanation dialogue model. We aimed to gain insights into three areas: 1. key components that make up an explanation interaction protocol (locutions); 2. relationships within those components (termination rules); and 3. component sequences and cycles (combination rules) that occur in explanations.

6.3.1 Design

We formulate our design based on an inductive approach. We use grounded theory [83] as the methodology to conceptualize and derive models of explanation. The key goal of using grounded theory, as opposed to using a hypothetico-deductive approach, is to formalize a model that is grounded on actual conversation data of various types, rather than a purely conceptual model.

The study is divided into three distinct stages, based on grounded theory. The first stage

consists of coding [83] and theorizing, where small chunks of data are taken, named and marked *manually* according to the concepts they might hold. For example, a segment of a paragraph in an interview transcript can be identified as an 'Explanation' and another segment can be identified as a 'Why question'. This process is repeated until the whole data set is coded. The second stage is categorizing, where similar codes and concepts are grouped together by identifying their relationship with each other. The third stage derives a theoretical model from the codes, categories and their relationship.

6.3.2 Data

We collected data from six different data sources encompassing six different types of explanation dialogues. Table 6.1 shows the explanation dialogue types, explanation dialogues that are in each type and the number of transcripts. Here, 'static' is defined as when an explainee or an explainer is the same person from transcript to transcript (e.g. same journalist interviewing different people). We gathered and coded a total of 398 explanation dialogues from all of the data sources. All the data sources[2] are text-based, where some of them are transcribed from voice and video-based interviews. Data sources consist of Human-Human conversations and Human-Agent conversations. We collected Human-Agent conversations to analyze if there are significant differences in the way humans carry out the explanation dialogue when they knew the interacting party was an agent with respect to the frequency of different locutions.

Table 6.1: Coded data description.

Explanation Dialogue Type	#Dialogue	#Scripts
1. Human-Human static explainee	88	2
2. Human-Human static explainer	30	3
3. Human-Explainer agent	68	4
4. Human-Explainee agent	17	1
5. Human-Human QnA	50	5
6. Human-Human multiple explainee	145	5

Data source selection was done to encompass different combinations of participant types and numbers. These combinations are given in Table 6.2. We diversify the dataset by including data

[2]Links to all data sources (including transcripts) can be found at https://explanationdialogs.azurewebsites.net

sources of different mediums such as verbal based and text-based.

Table 6.2: Explanation dialogue type description.

Participants	Number	Medium	Data source
1. Human-Human	1-1	Verbal	Journalist Interview transcripts
2. Human-Human	1-1	Verbal	Journalist Interview transcripts
3. Human-Agent	1-1	Text	Chatbot conversation transcripts
4. Human-Agent	1-1	Text	Chatbot conversation transcripts
5. Human-Human	n-m	Text	Reddit AMA records
6. Human-Human	1-n	Verbal	Supreme court transcripts

Table 6.3 presents the codes and their definitions. We identify 'why', 'how' and 'what' questions as questions that ask counterfactual explanations, questions that ask explanations of causal chains, and questions that ask causality explanations respectively. The whole number of the code column refers to the categories the codes belong to, where 1) Dialogue boundary; 2) Question type; 3) Explanation; 4) Argumentation; 5) Return question type.

6.4 Grounded Explanation Interaction Protocol

In this section, we present the interaction model resulting from our grounded study, and formalize the model using agent dialogue framework (ADF) [150] as an atomic dialogue type [240]. Note that a dialogue can range from a purely visual user interface interaction to verbal interactions. We analyse some observed patterns of interaction and compare the grounded model to an existing conceptual model.

When formalizing the model, we consider the interaction between explainer and explainee as a dialogue game. Dialogue games are depicted as interactions between two or more players. The

Table 6.3: Code description.

Code	Description
1.1 QE start	Explanation dialogue start
1.2 QE end	Explanation dialogue end
2.1 How	How questions
2.2 Why	Why questions
2.3 What	What questions
3.1 Explanation	Explanation given for questions
3.2 Explainee Affirmation	Explainee acknowledges explanation
3.3 Explainer Affirmation	Explainer acknowledges explainee's acknowledgment
3.4 Question context	Background to the question provided by the explainee
3.5 Counterfactual case	Counterfactual case of the how/why question
4.1 Argument	Argument presented by explainee or explainer
4.2 Argument-s	An argument that starts the dialogue
4.3 Argument-a	Argument Affirmation by explainee or explainer
4.4 Argument-c	Counter argument
4.5 Argument-contrast case	Argumentation contrast case
5.1 Explainer Return question	Clarification question by explainer
5.2 Explainee Return question	Follow up question asked by explainee

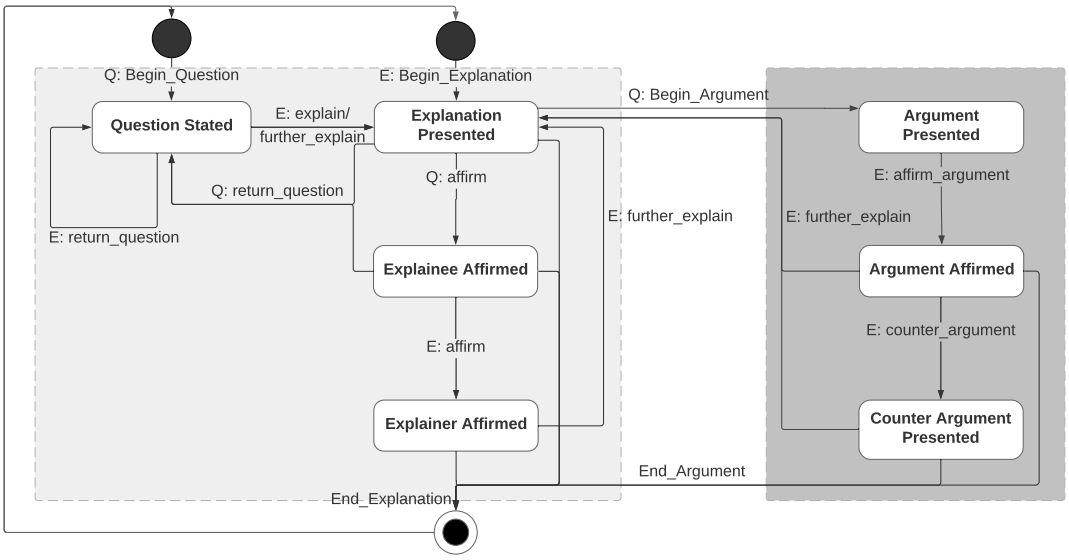


Figure 6.1: Explanation Dialogue Model

players can make 'moves' with utterances, according to a set of rules. Dialogue game models have been used to model human-computer interaction [20], to model human reasoning [184] and to develop protocols for interactions between agents [61].

Formal dialogue models have been proposed for different dialogue types [240], such as negotiation dialogues [8], persuasion dialogues [240] and a combination of negotiation and persuasion dialogues [61]. To the best of our knowledge there is no formal explanation dialogue game model grounded on data.

In essence, Figure 6.1 depicts the state model resulting from the human explanation analysis. At a glance, there are two sub-dialogues happening inside larger explanation dialogue; explanation and argumentation. This model capture the dialogue acts, beginning and ends of the dialogues and dialogues transitions. In the sections below, we explain the model in detail.

6.4.1 Agent Dialogue Framework

We use McBurney and Parson's agent dialogue framework [150] to formalize the explanation dialogue model as a dialogue game. The Agent Dialogue Framework (ADF) provides a modular and unifying framework that can represent and combine different types of atomic dialogues in the typology of Walton and Krabbe [240], with the freedom of introducing new dialogue type combinations. The ADF has three layers: 1. topic layer; 2. dialogue layer; and 3. control layer. In the topic layer, the topics of discussion in a dialogue game are presented in a logical language. Then, the dialogue layer [150] consists of a set of rules:

Commencement rules: rules under which the dialogue commences.

Locutions: Rules that determine which utterances are permitted in the dialogue-game. Typical locutions include assertions, questions, arguments, etc.

Combination rules: Rules that define the dialogical context of the applicability of locutions. E.g. it might not be applicable to assert preposition p and $\neg p$ in the same dialogue.

Commitments: Rules that determine the circumstances where players express commitments to a preposition.

Termination rules: Rules that determine the ending of a dialogue.

More formally, given a set of participating agents $\mathcal{A}$, we define the dialogue G at the dialogue layer as a 4-tuple $(\Theta, \mathcal{R}, \mathcal{T}, \mathcal{CF})$, where Θ denotes set of legal locutions, $\mathcal{R}$ the set of combinations, $\mathcal{T}$ the set of termination rules and $\mathcal{CF}$ the set of commitment functions respectively [150].

Selection and transitions between dialogue types are handled in the control layer. Dialogue types can be combined using iteration, sequencing and embedding [150, pg 10]. When combined, an ADF is given by 5-tuple $(\mathcal{A}, \mathcal{L}, \Pi_a, \Pi_c, \Pi)$ where the set of agents is given by $\mathcal{A}$, logical language representation given by $\mathcal{L}$, set of atomic dialogue types given by Π_a, set of control dialogues given by Π_c, and Π is the closure of $\Pi_a \cup \Pi_c$, which represents the set of formal dialogues denoted by the 4-tuple given above. Closure is defined under the combination rules presented by McBurney and Parsons [150].

6.4.2 Formal Explanation Dialogue Game Model

In this section we present the formal explanation dialogue model as a new atomic dialogue type [240] using the modular ADF, and discuss how it is derived from the grounded data of explanation dialogues according to the layers of ADF. Our analysis of the data shows that people switch from explanation to argumentation and back again during an explanation dialogue, in which the explainee questions a claim made by an explainer. For this reason, our model has two dialogue types: Explanation and Argumentation. Dialogue games of atomic dialogue types [240] have an initial situation and an aim (e.g persuasion dialogue having the initial situation of conflicting opinions of the interacting party and the aim of resolving the conflict). For our explanation dialogue, the initial condition is the knowledge discrepancy between explainer and explainee of the topic p and the aim is to provide knowledge about the topic p to the explainee.

Formally, the explanation dialogue model (ADF_E) is the tuple:

$$ADF_E = (\mathcal{A}, \mathcal{L}, \Pi_a, \Pi_c, \Pi) \tag{6.1}$$

where the set of agents $\mathcal{A} = \{Q, E\}$, where labels Q and E refer to the Questioner (the explainee) and the Explainer respectively; $\mathcal{L}$ is the set of logical representations about topics (denoted by p, q, r, ...), $\Pi_a = \{G_E, G_A\}$, where G_E is the explanation dialogue and G_A is the argumentation dialogue, Π_c = (Begin_Question, Begin_Explanation, Begin_Argument, End_Explanation, End_Argument), and Π is the closure of $\Pi_a \cup \Pi_c$ under the combination rule

set. Π gives us the set of formal explanation dialogue G.

The **Topic Layer** is dependent on the particular application domain in which the explanation dialogue is embedded, so we do not define this further.

Dialogue Layer The dialogue layer consists of the two dialogue types: explanation (G_E) and argumentation (G_A):

$$\begin{aligned} G_E &= (\Theta_E, \mathcal{R}_E, \mathcal{T}_E, \mathcal{CF}_E) \\ G_A &= (\Theta_A, \mathcal{R}_A, \mathcal{T}_A, \mathcal{CF}_A) \end{aligned} \tag{6.2}$$

The set of legal locutions are defined by:

$$\begin{aligned} \Theta_E &= (explain, affirm, further\_explain, return\_question) \\ \Theta_A &= (affirm\_argument, counter\_argument, further\_explain). \end{aligned} \tag{6.3}$$

For clarity, we define the commencement rules, combination rules, and termination rules via the state transition diagram in Figure 6.1. While most codes are directly transferred to the model as states and state transitions, codes that belonged to the information category are embedded in different states. The combination rules $\mathcal{R}_E$ and $\mathcal{R}_A$ are defined by the individual transitions on the diagram. For example, after a dialogue begins with a question, the next locution is either the explainer asking for clarification using a return_question or giving an explanation. Similarly, the set of termination rules can be extracted from the state model as the state transitions that lead to the termination state, giving $\mathcal{T}_E$ = (affirm(p), explain(p)) and $\mathcal{T}_A$ = (affirm_argument(p), counter_argument(p)). We do not define commitments $\mathcal{CF}$ as these were not observable in our data.

Control layer This can be identified as state transitions that lead to and out of the two dialogue types in Figure 6.1 (e.g. argue, explanation_end). Argumentation occurs naturally within explanation dialogues, meaning that this is an *embedded* dialogue, as defined by McBurney and Parsons [150]. An argument can occur after an explanation was given, which will then continue on to an argumentation dialogue. The dialogue then returns to the explanation dialogue, as shown in Figure 6.1. A single explanation dialogue can contain many embedded argumentation

dialogues.

Explanation dialogues can occur in sequence, which is modelled by the external loop. Note that a loop within the explanation dialogue implies that the ongoing explanation is related to the same original question and topic, while a loop outside of the dialogue means a new topic is introduced. We coded explanation dialogues to end when a new topic was raised in a question. Questions that ask for follow-up explanations (return_question) were coded when the questions were clearly identifiable as requesting more information about the given explanation.

Example: We now go through the formal model with an example dialogue which is taken from the human-agent experiments discussed in Section 6.5. Example is given in Table 6.4 with the dialogue text, locutions/rules and a commentary about the dialogue. Two agents who are explainee (player) and the explainer (agent) participate in the dialogue given by Q and E respectively and the topic 'cities' by p:

Table 6.4: Example: from human-agent experiments of Ticket to Ride domain.

Dialogue Text	Locutions/Rules	Commentary
E**: Opponent is gazing at Duluth to Omaha route and will try to extend it to Kansas City.**	*Begin_Explanation(p)*	- Commence explanation dialogue with an explanation about cities which the opponent is gazing at p.
Q**: Is he going to Pittsburgh?**	*return_question(p)*	- Using locution *return_question* available in Explanation dialogue type, inquiring more information
E**: Opponent will try to Extend the path from Pittsburgh to Houston through Atlanta, has been repeatedly gazing at that path**	*further_explain(p)*	- providing further explanation using *further_explain* locution about topic p.
Q**: No. He is going to El Paso.**	*Begin_Argument(p)*	- Argumentation sub-dilaog begins about topic p after.
E**: Yes, now opponents gaze is focused at El Paso and will try to build from Little rock to Dallas to El Paso.**	*affirm_argument(p)*	- Argument is acknowledged by the Agent (E) using locution *affirm_argument*
.	*End_Argument(p)*	- End the embedded argument dialogue by using the control dialogue *End_Argument* which also ends the initial dialogue.

This example shows an interaction between an agent and a human using the explanation dialogue

with an embedded argumentation dialogue. The human-agent study is discussed in depth in Section 6.5. The example demonstrates the ability of our model to handle embedded dialogues and cyclic dialogues (explanation dialogues that occurs twice with Begin_explanation and further_explain) which is a similar model of explanation dialogue by Walton [25] lack. Detailed model comparison between our model and Walton's can be found in Section 6.4.4.

6.4.3 Analysis

We focus our analysis on three areas to further reinforce the derived interaction protocol: 1. Key components of an Explanation Dialogue; 2. Relationships between these components and their variations between different dialogue types; and 3. The sequence of components that can successfully carry out an explanation dialogue.

Code Frequency Analysis

The average code occurrence per dialogue in different dialogue types is depicted in Figure 6.2. In all dialogue types, a dialogue is most likely to have multiple *what* questions, multiple *explanations* and multiple *affirmations*.

Argumentation is a key component of an explanation dialogue. The explainee can have different or contrasting views to the explainer regarding the explanation, at which point an argument can be put forth by the explainee. An argument in the form of an explanation that is not in response to a question can also occur at the very beginning of an explanation dialogue, where the argument set the premise for the rest of the dialogue. An argument is typically followed by an affirmation and may include a counter argument by the opposing party. From Figure 6.2, Human-Human dialogues with the exception of QnA have argumentation but Human-Agent dialogues lack any substantial occurrences of argumentation.

Explanation Dialogue Termination Rule Analysis

Participants should be able to identify when a dialogue ends. We analyse the different types of explanation dialogues to identify the codes that are most likely to signify termination.

From Figure 6.3, all explanation dialogue types except Human-Human QnA type are most likely to

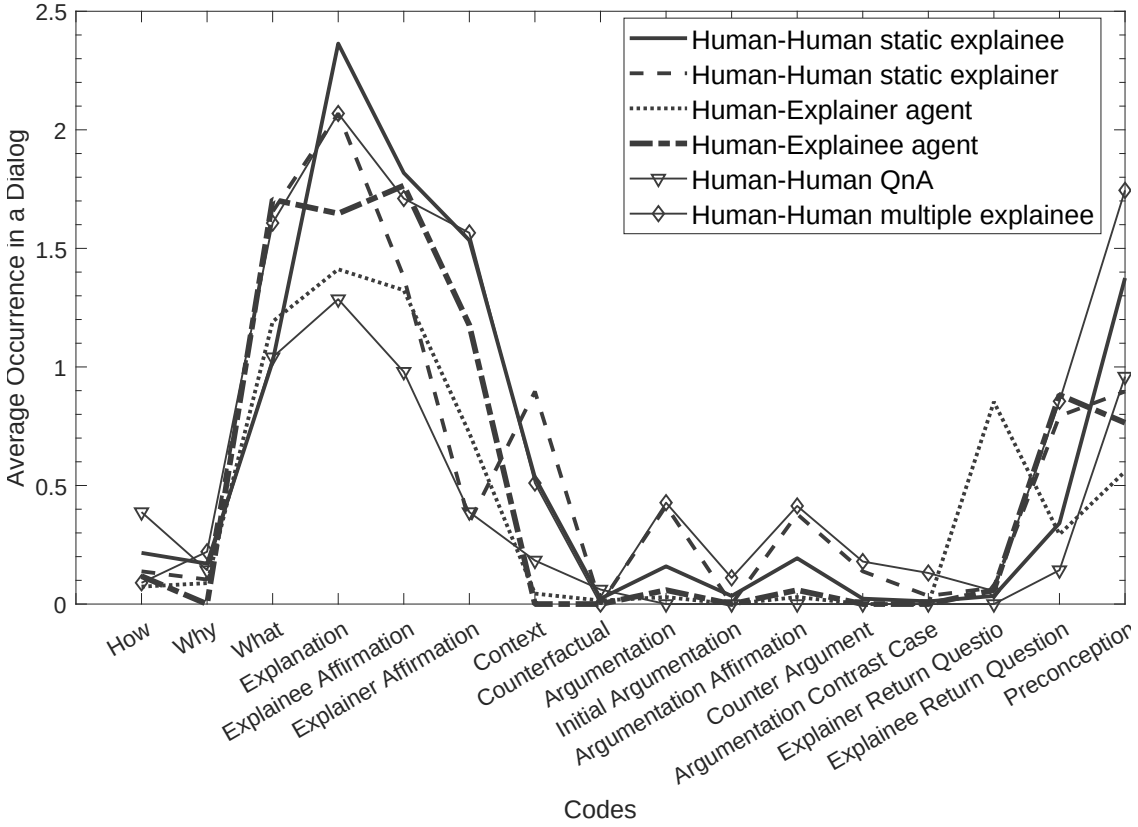


Figure 6.2: Average code occurrence per dialogue in different explanation dialogue types

end in an explanation. The second most likely code to end an explanation is explainer affirmation. Ending with other codes such as explainee and explainer return questions is presented by 'Dialogue ending In Other' bar in Figure 6.3. It is important to note that although a dialogue is likely to end in an explanation, that dialogue can have previous explainee affirmations and explainer affirmations.

6.4.4 Model Comparison

We compare the explanation dialogue model, which also contains an argumentation sub-dialogue, by Walton [25]. Walton proposed the model shown in Figure 6.4, which consists of 10 components. This model focus on combining explanation and examination dialogues with argumentation. A similar shift between explanation and argumentation/examination can be seen between our model and Walton's. According to the data sources, argumentation is a frequently present

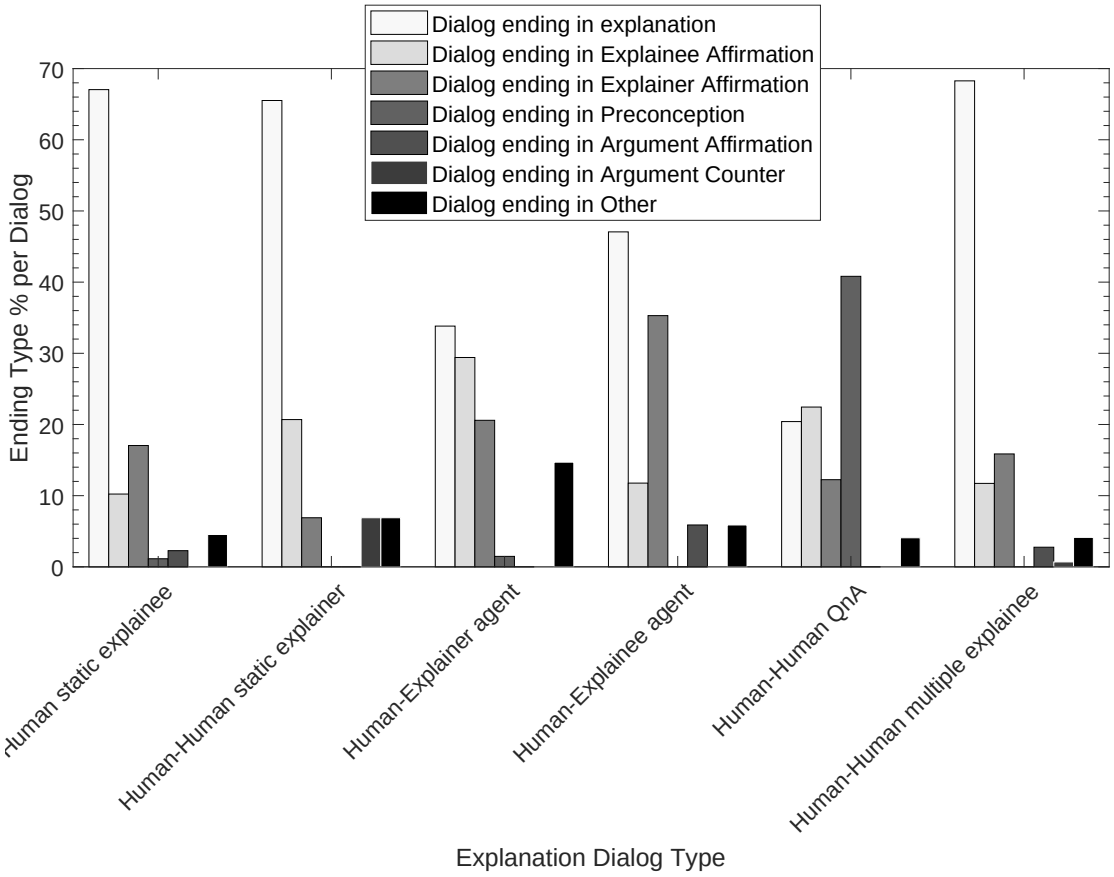


Figure 6.3: Average code occurrence in per dialogue in different explanation dialogue types

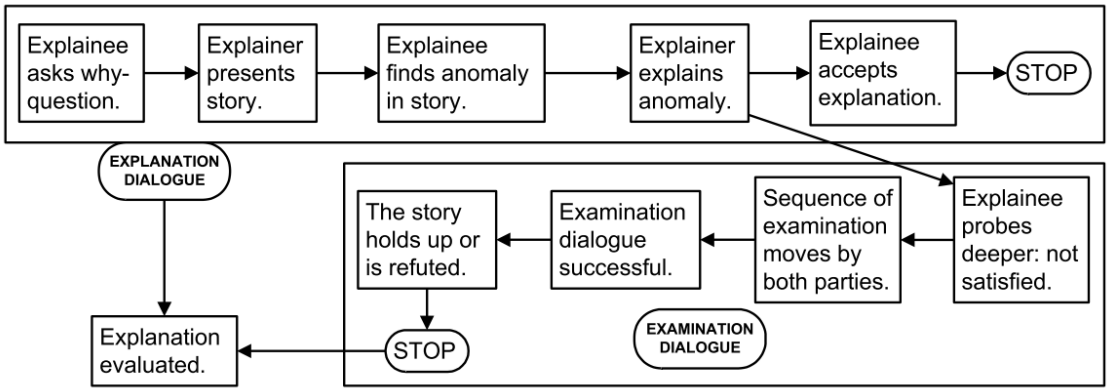


Figure 6.4: Argumentation and explanation in dialogue [25]

component of an explanation dialogue, which is depicted by the Explainee probing component in Walton's Model. The basic flow of explanation is the same between the two models, but the

models differ in two key ways. First, is the lack of *examination* dialogue shift in our model. Although we did not derive an examination dialogue, a similar shift of dialogue can be seen with respect to affirmation states. That is, our 'examination' is simply the explainee affirming that they have understood the explanation. Second is Walton's focus on the evaluation of the successfulness of an explanation in the form of examination dialogue, whereas our model focuses on delivering an explanation in a natural sequence without an explicit form of explanation evaluation.

Thus, we can see similarities between Walton's conceptual model (Figure 6.4 and our data-driven model (Figure 6.1). The differences between the two are at a more detailed level than at the high-level, and we attribute these differences to the grounded nature of our study. While Walton proposes an idealised model of explanation, we assert that our model captures the subtleties that would be required to build a natural dialogue for human-agent explanation.

6.5 Empirical Validation

In this section, we discuss the validation of the derived explanation interaction protocol. We conducted a human-agent study in which an agent provides explanations using our model. The purpose of the study is to test whether the proposed model holds in a human-agent setting, and in particular, that the human participants follow the dialogue model when interacting with an artificial agent. The ethics approval to conduct the study was provided by The University of Melbourne's human ethics committee (ID: 1647972).

6.5.1 Study

We conducted our study using the *Ticket to Ride*[3] online computer game in a co-located competitive setting, previously used by Newn et al. and Singh et al., in a university usability lab. The basic layout of the game is shown in Figure 6.5. In this game, players must compete to build train routes between two cities, with each player building at least two such routes. For the purpose of the study, a game is played between two players who we term as the *player* and the *opponent*. The *player* is assisted by an intelligent software agent that predicts the intentions and plans of the opponent. It is important to note that for a two player game, each route can be claimed only by

[3]https://www.daysofwonder.com/tickettoride/en/

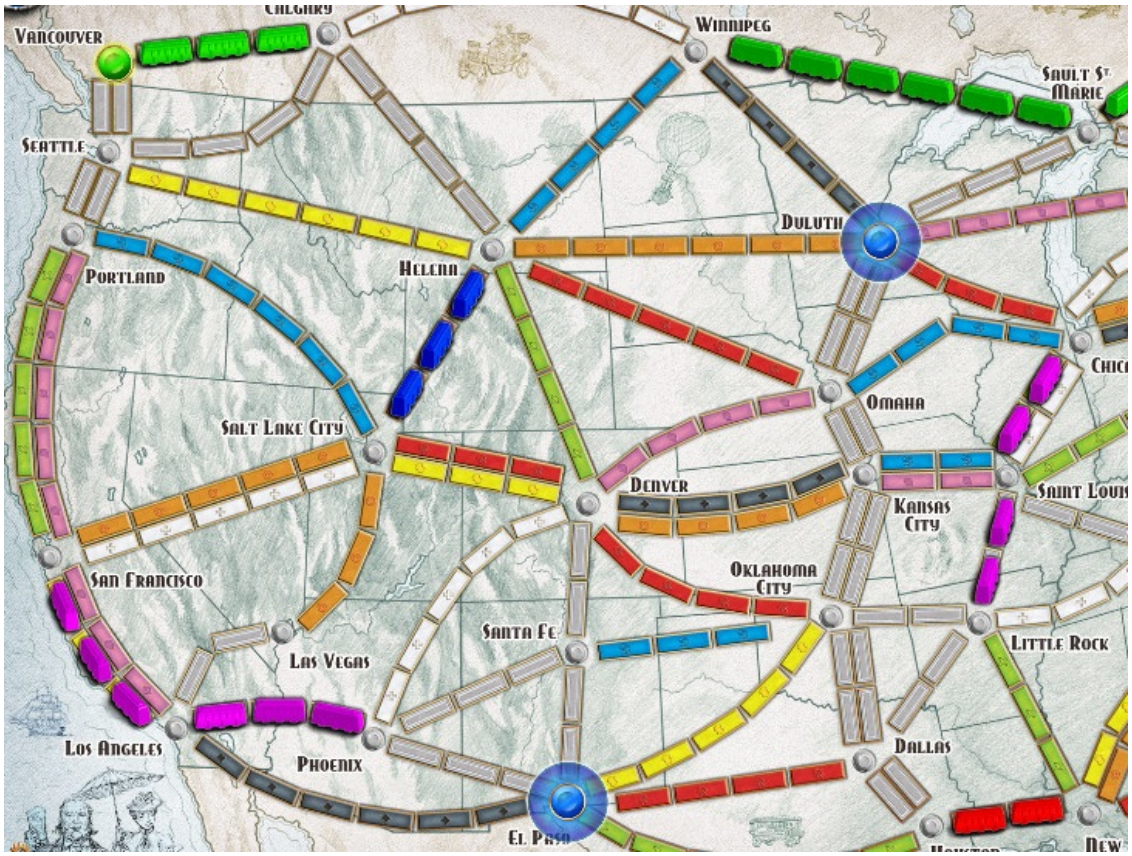


Figure 6.5: Ticket to Ride Computer Game

one player. This allows players to block each other deliberately or otherwise, therefore inferring the intent of the opponent is beneficial for winning the game. We use the intent recognition algorithm of Singh et al. to predict the opponent's future moves. The algorithm uses gaze data from an eye tracker and the actions of the opponent to formulate the possible plans (e.g. most probable routes the opponent can take). The opponent's gaze will also appear as a heat map on top of the player's Ticket to Ride game screen[4]. The agent communicates the predictions and their explanations to the player through a chat window.

To evaluate our model, we adopted a *Wizard of OZ* approach described by Dahlbäck, Jönsson, and Ahrenberg, meaning that the natural language generation of the explanation agent is played by a human 'wizard', but this is unknown to the human participant. It is important to note

[4]Video capture of an experiment is provided in supplementary material at https://explanationdialogs.azurewebsites.net/supp.zip

that only the natural language generation is delegated to the wizard, while the agent generates and visualize plans (a set of connected train routes) in a separate interface in order to assist the wizard. Wizard has access to the visualized plans, most gazed at routes and most gazed at cities. The argument for using the Wizard of Oz (WoZ) technique as opposed to a natural language implementation is twofold. First, to gather high quality empirical data related to the model bypassing the limitation that exists in natural language interfaces [54]. Second, having an interaction that closely resembles human discourse [54] allows the human to have more natural responses. Wizard of Oz techniques have been demonstratively shown to successfully evaluate conversational agents [47, 66], human-robot interactions [18] and automated vehicle-human interactions [147, 174].

The *Wizard* uses the prediction information of the agent and translates it to a more natural dialogue, enabling us to get empirical data on natural explanation dialogues between the player and the agent. The player is informed that he/she can communicate with the agent in natural discourse. A prediction of the game includes a route that the opponent might build (e.g from Pittsburgh to Houston) and a city area opponent might be interested in (e.g Interested around Houston). Predictions are generated from the implemented intent recognition algorithm of Singh et al. The wizard follows a simple natural language template: prediction followed by the explanation. The explanation template can include one or more of the following in any order: gaze explanation (e.g the opponent has been repeatedly gazing at that path) and causal history explanation (e.g the opponent has already built some paths along that route).

The protocol of the Wizard is outlined as follows. The Wizard follows the locutions, termination rules and combination rules of the dialogue model. Predictions and explanations of the agent is translated to natural language using the template described above by the wizard according to the nature of the locution used. Players (experiment participants) can ask questions and present arguments in natural language. Players can initiate dialogues as well as reply to the Wizard at any time in the game, and are in control of the frequency of the interaction. The wizard follows a static failure response to any interactions that failed or is unable to provide predictions and explanations (e.g I'm unable to answer that). Note that in the case of a participant using an invalid locution or a control dialogue, the dialogue will fail and the wizard will end the dialogue with a termination rule.

The parameters of the experiments are as follows. In total, we obtained 101 explanation dialogues across 14 experiments. Players were from the same university, aged between 23 and 31 years ($M = 27.2$). Players were observed through an observation room, in which the wizard was located. The duration of each experiment had an upper bound of 30 minutes ($M = 20.15$), limited by the duration of the game-play, with the ability to end early if the game is won by either side before 30 minutes (game ends when all trains have been used by either side). During game-play the player has the freedom to engage in conversation with the agent or play the game disregarding the agent, thus each experiment yields a different number of dialogues ($M = 7.21$). Conversations between the player and the wizard were carried out using a chat client, through which we recorded the dialogue data. Extracted data were then analysed according to locutions, control dialogues and their sequences.

6.5.2 Results and Findings

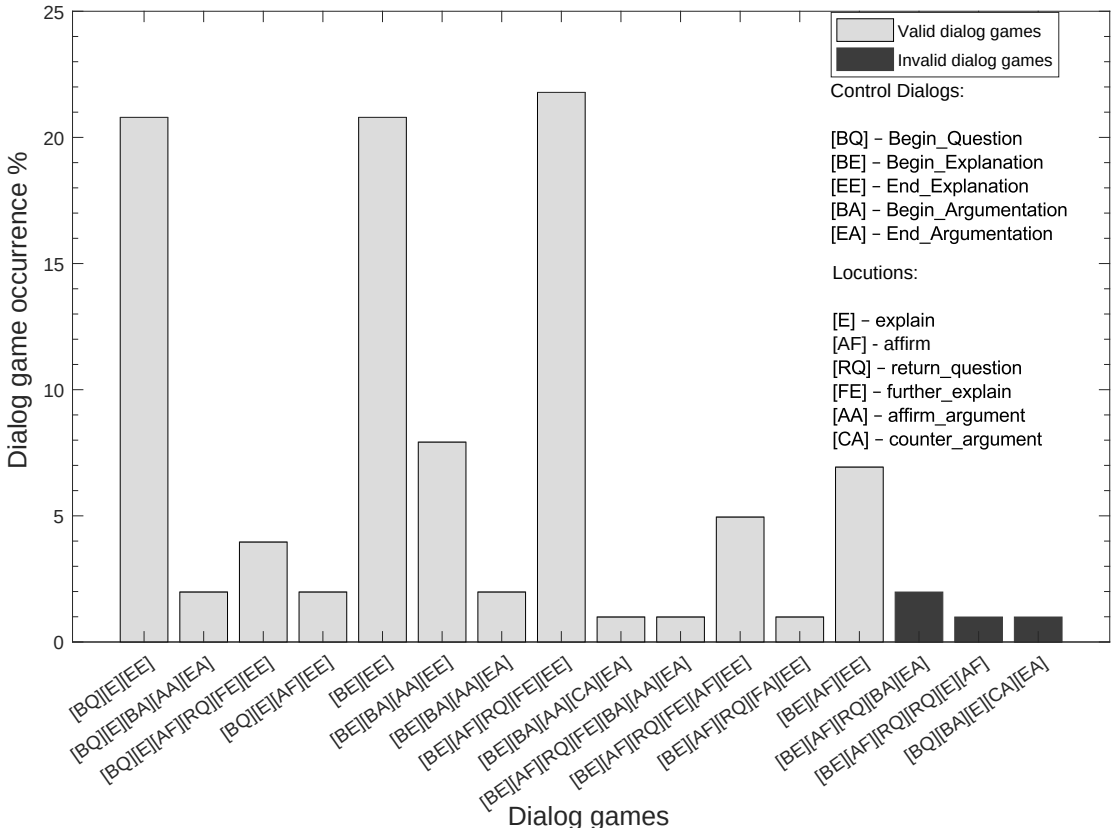


Figure 6.6: Empirical results of human-agent dialogue games.

Figure 6.6 illustrates the explanation dialogue games in the human-agent study. Control dialogues and locutions are depicted by shortened tags, which forms a sequence when combined. An example of a dialogue game using the tags would be [BQ][E][AF], which corresponds to [*Begin_question* ⇒ *explain* ⇒ *affirm*] in locutions and control dialogues. Figure 6.6 shows percentages a specific dialogue game type occurred, and invalid dialogue games.

The proposed explanation dialogue model held true for 96 out of 101 dialogue game instances we observed. Figure 6.6 indicates the 5 invalid dialogue game types that occurred. These dialogues became invalid dialogue games according to our model because of the parallelization of dialogue combination moves. For example, consider the dialogue game [BE][AF][RQ][BA][EA]. Here, after affirm and *return_question* locution, *Begin_Argumentaion* control dialogue occurs. This sequence is illegal according to the model. If parallelization is allowed, *Begin_Argumentaion* control dialogue can occur without waiting for a termination rule (e.g. affirm, explain). We attribute this limitation to the nature of the grounded data where parallelization cannot be accurately captured. This limitation can potentially be rectified by introducing parallelization [150] to the combination rules.

6.6 Conclusion

Explainable Artificial Intelligent systems can benefit from having a proper interaction protocol that can explain their actions and behaviours to the interacting users. Explanation naturally occurs as a continuous and iterative socio-cognitive process that involves two (sub)processes: a cognitive process and a social process. Most prior work is focused on providing explanations without sufficient attention to the needs of the explainee, which reduces the usefulness of the explanation to the end-user.

This chapter proposes a interaction protocol for the socio-cognitive process of explanation that is derived from different types of natural conversations between humans as well as humans and agents. We formalise the model using the Agent Dialogue Framework [150] as a new atomic dialogue type [240] of explanation with an embedded argumentation dialogue, and we analyse the frequency of occurrences of patterns. To empirically validate our model, we undertook a human behavioural experiment involving 14 participants and a total of 101 explanation dialogues. Results indicate that our explanation dialogue model can closely follow Human-Agent explanation

dialogues. The main contribution of this chapter lies in the formalized interaction protocol for explanation dialogues that is grounded on data, a secondary contribution is the coded (tagged) explanation dialogue data-set of 398 dialogues. By following a data-driven approach, the proposed model captures the structure and the sequence of an explanation dialogue more accurately and allow natural interactions than explanations from existing models. The main contribution of this chapter is a grounded interaction protocol derived from explanation dialogues, formalized as a new atomic dialogue type [240] in the ADF. XAI systems that deal in explanation and trust will benefit from such a model in providing better, more intuitive and interactive explanations.

There are several future directions that this work can evolve into in the context of interactive explanations. Other forms of interaction modes can be introduced such as visual interactions which may include different forms of combination and termination rules. Recent work has also taken influence of interactive explanations proposed here to introduced sequential explanations for RL agents [251]. Explanation dialogue models can be used as the explanation interface to the underlying explainable model, e.g. the work discussed in Chapters 3 and 4 can be instantiated as the explainable model that underpins the explanation dialogue model.

Chapter 7

Conclusion

This thesis presents original work that contributes to the body of explainable reinforcement learning literature. We proposed and developed computational model-agnostic explainable models for reinforcement learning agents, that can generate local explanations about agent's actions for 'why' and 'why not' questions. These models were largely inspired by the theories of cognitive science relating to causal explanations. Further enhancements were done for the causal explanation model through the insights gained via human explanations. Specifically, the primary outcome of the thesis is the introduction of the *action influence model* and the *distal explanation model*. Original research was done to learn action influence models end-to-end, by developing action influence discovery algorithms. This thesis further proposes an explanation dialogue model that can facilitate interactive explanations and can function as the explanation interface between the explainable RL model and the explainee.

The majority of the work described in this thesis followed a human-centred approach. That is, in generating, selecting, structuring and communicating the explanations, concepts and theories were build upon the prevalent literature in cognitive and social sciences. A human-centred approach can plausibly provide 'better' explanations to the end-users and the empirical results reported in this thesis support this. This can also serve as a motivation for future research in following a human-centred approach for explainability. This Chapter organises the research contributions, insights and findings obtained by answering the main research questions given in Chapter 1. A commentary of the primary contributions and the developed artefacts is done, giving limitations of each. Limitations of the explainable RL models, dialogical explanations and challenges in combining them as a whole is discussed giving concrete future directions.

7.1 Research Contribution

This thesis addressed four main research questions, and proposed new models and definitions for explainable reinforcement learning agents. This section reflects upon the specific research contributions and artefacts produced by answering the research questions and how they are situated in the literature. Table 7.1 summarises the research contribution of this thesis.

Research Question	Aims	Contribution
RQ1: How can reinforcement learning agents provide *causal* explanations of agent behaviour that increase the understanding and trust of the users?	• Develop a human-centred explainable model for RL agents. • Propose definitions and formalisations to generate explanations for 'why' and 'why not' questions of agent actions.	• The Action influence model, as a model-agnostic explainable model for RL agents.
RQ2: Do *distal* causal explanations of RL agents improve the intelligibility of the agent behaviour?	• Develop a explainable model that is grounded on human explanation data for RL agents. • Propose new definitions and formalisations to generate explanations. • Improve upon the action influence model accuracy.	• The Distal explanation model. • Coded human explanations of RL agents in StarCraft II domain.
RQ3: How can reinforcement learning agents *discover the causal action influence structure* of the domain?	• Learn the action influence structure from RL agents' interaction data.	• The action influence learning architecture and models.
RQ4: What are the patterns of dialogical explanation that can facilitate *interactive explanations* in XAI systems?	• Develop a human-centred general explanation dialogue model for explainable agents. • Propose definitions and formalisations of the explanation sequence structure.	• The Dialogue explanation state model, defined by the agent dialogue games. • Coded corpus of human explanation dialogues.

Table 7.1: Research contribution, proposed models, definitions and artifacts.

7.2 Causal Explanations in Explainable Reinforcement Learning

A major driver when developing an explainable system is the *type* of the explanation that is presented to the explainee. The type of the explanation affects the generation, selection, contrasting and structure of the explanation. Causal explanations are one such type of explanation. To our knowledge, the Action influence model is the first such model that is made to generate causal explanations for reinforcement learning agents, answering **RQ1**.

Motivation for exploring causal explanations of RL agents has clear roots in cognitive science, philosophy and social science literature as elaborated in Chapters 2 and 3. People in general have a strong affinity to causal explanations in comparisons to other types (state-action explanations, visual explanations) and indeed this was observed in the empirical evaluation discussed in Chapter 3. Further, causal explainability and interpretability methods have found various levels of success in supervised learning [164], paving a new avenue for causality based explainability research.

The type of questions the explainee can query the agent can also vary, from 'What', 'How', 'Why' and 'Why not'. RQ1 was focused on developing explanation generation methods for 'Why' and 'Why not' questions. This was motivated by the work of Penney et al. [179], that note the most frequent types of questions asked from agents were 'Why' and 'Why not' questions. Though the action influence model explicitly focused on generating explanations for those types, our model can be extended to answer 'What' and 'How' questions without fundamental changes. This is due to causal models being in the top ring of the causal ladder [178], enabling them to answer questions on associative (What) and interventionist (How) questions.

In answering RQ1, we considered explanation generation for a question posed for the agent's action, which is a form of local explainability. Other works have found success in explaining the behaviour (represented by the policy of the agent) [99] and summarising the behaviour as explanations [9]. Though global explanation methods provide an overview of the agent intelligibility, we believe local explanations are also needed for a granular and a more nuanced understanding of the agent. The need for this understanding can be visible in human-collaboration scenarios (discussed below in RQ2). A hybrid approach can also plausibly perform better, and the action influence models can serve as the basis to generate causal policy explanations of RL agents.

The action influence model is formalised using structural causal models [92], augmented with

actions based on their causal influence. This formalism allowed the model to handle 'Why not' questions by simulating counterfactuals and generating contrastive explanations. While there are different ways to generate contrastive explanations [236, 122, 155], our approach that contrast the *causal chains* can provide causal information that earlier methods lack. The explanation *selection* is another variable that can change the explanation product. In Chapter 3, we defined a minimally complete explanation for action influence models, which included the reward nodes, immediate nodes (before the reward nodes) and the header node of the action. This definition was motivated by the need to include both the long term and the short term motivation of the agent in the explanation. Other researchers can choose to use different proxies and heuristics to define minimality such as: selecting the highest impacted node in the causal chain, choosing the node based on the explainee's epistemic state. The action influence models are flexible to allow such different selection methods, and this presents a possible direction in extending these models.

The effectiveness of causal explanations generated using action influence models was assessed in both computational and human experiments, which showed significant improvements compared to other baseline local explanation models (state-action, descriptive). Limitations of the evaluation phase are discussed in the sections below. Action Influence models were targeted at reinforcement learning agents and were specifically evaluated in model-free RL agents. As these models act as surrogate models, most RL agent architectures are supported, given that the replay data of the agent is accessible. Importantly, due to its surrogate nature, action influence models can be extended to other sequential decision making agents (e.g. planning agents) with states, actions and rewards.

7.3 Distal Explanations

Humans often expect familiar modes of communications, and explainable systems need to adhere to these needs [131]. A human grounded approach for explainability can have the potential to be more effective and accepted by the end-users. Keeping this as the motivation, in **RQ2**, we followed a data-driven grounded approach to develop an explainable model for RL agents. A distal explanation model is proposed as the answer to the RQ1.

We obtained data from 30 participants in a human-agent experiment, where the participants

formulated their own explanations of StarCraft II RL agents after seeing the agents in action. Importantly, participants had no restrictions on the format and the structure of the explanation, and used free text as the medium. Thematic analysis was used to conceptualise the 240 explanations that were obtained into distinct codes. We observed that the majority of the participants referred to the agents' state features, actions and causal relationships in the explanations. Intuitively this made sense, as these codes convey the bulk of the knowledge that is needed for intelligibility. Participants also noted some *temporal* dependency of the agent's action, which prompted the development of a distal explanation model. This temporal causal nature is described as the opportunity chains, and explain the causal dependencies an action can have on future actions of the agent.

RQ2, in addition to introducing the distal form of explanation, also improved the accuracy of the action influence models. In the base action influence model, the causal effects were estimated using structural causal equations. Though the accuracy of the task prediction using these functions was acceptable to generate explanations, further improvements were possible. In lieu of structural equations, we used decision trees to approximate the RL agent's policy. Action influence models were used to extract the causal chains and a recurrent neural network was used to predict the distal action of the agent. These three sub-models formed the distal explanation model that addressed the RQ3. Other work have also used decision trees as surrogate models for explainability [26, 206, 62], though it is possible to use other forms of RL agent policy approximation methods [233, 112, 231, 33]. Further, to predict the distal action, other architectures can be used such as action consequence predictions [42] and sequential action predictions [154].

Most domains that exist for RL agent evaluations can be ill-suited to test the strength of an explainable model. StarCraft II somewhat alleviate this challenge due to its large distinct action and state space. Indeed, explainability researchers have proposed StarCraft II as an explainability benchmark for RL agents [181]. In Chapter 4, we used the StarCraft II framework to develop new scenarios focused on adversarial, search and rescue and collaborative tasks. In particular, the collaborative task allowed the participants of the experiment to interact with the agent to complete a task, where the distal explanation model performed significantly better. The usefulness of distal explanations can have implications for other types of scenarios like the join

task planning and the human in the loop RL agents.

7.4 Learning the Action Influence Structure

Action influence models and distal explanation models rely on the pre-defined causal influence structure of the agent. In the work presented in Chapters 3 and 4, this causal influence structure is handcrafted through the examination of the constraints set in the RL environment. Though this is feasible in small domains and simulations, for real-world problems with large state and action spaces, handcrafting the influence structure becomes more difficult. **RQ3** seeks to answer this challenge through learning the causal action influence structure from the RL agent's interaction data (state-action traces) itself, without the need of a domain expert.

RQ3 specifically focuses on discovering the causal action influences an RL agent might have within its state variables. The important distinction between the action influence discovery and the traditional problem of causal discovery lies in how the former also learns the *action* that influenced the causal relationship between some state variables, in addition to learning the causal links between variables. In essence, in Chapter 5, we develop algorithms and models that can learn the structure of the action influence models. When the structure is present, structural equations can be inferred and trained as needed for the action influence models, or can be used to extract the causal chains for distal explanation models.

Chapter 5 introduces methods to capture the influences of actions, by encoding them in accordance with the state variable changes of the next instance. By using a data-set of such encoded actions (an action associated with one or more state variables), the problem of action influence discovery can be cast to a causal discovery algorithm. The encoded influence is treated as another variable in the causal discovery process. We used Zhu, Ng, and Chen [257]'s causal discovery method for this process, though using other traditional score-based and constraint-based causal discovery algorithms is also possible. The resulting causal graph from the discovery process does not have actions annotated to the edges of the graph. This graph can be transformed into an action influence graph by decoding the nodes that have actions encoded into them by merging and adding incoming edges that are action annotated. The accuracy of the architecture was evaluated in RL domains by comparing the difference the learned model had against the ground truth action influence graph.

Implications of the findings in Chapter 5 affect not just the explainability but the learning mechanisms of RL agents. The learning of the action influence structure happens throughout the agents' lifespan, that is the data from all the episodes of an agent is used, not just the data from an optimal policy that is converged. The sub-optimal state-action data from the exploration is also captured here, which allows the discovery of infrequently observed causal connections to be formed. In explainability, this is vital when generating counterfactual contrastive explanations, to justify why a specific action was not taken. Having an automated learning mechanism for action influence can also help to guide the RL agent's learning process when integrated into the exploration of the agent.

7.5 Dialogical Explanations

Chapters 3 to 5 focused solely on proposing explainable models for RL agents. In Chapter 6, we focused on the explanation interface that explainable models need to communicate the explanations. **RQ4** sought to understand the nature and the structure an explanation dialogue can have and how to formalise this dialogue to build computational models. To answer RQ4, Chapter 6 introduced a data-driven state model that captures the explanation dialogue between the explainer and the explainee, grounded on human explanation dialogues.

RQ4 presented the challenge of having a deeper understanding of the structure and a sequence of human explanation dialogue as a prerequisite to building a state model an agent can use in the explanation interface. To this end, we gathered 398 human explanation dialogue transcripts from different sources like journalist interviews and Reddit treads. When grounding computational models on human data, it is important to capture the two-way interactions of different types of explainers and explainees. Thus, a variety of human-human and human-agent dialogues were selected, where the human and agent played the role of explainer and the explainee interchangeably. This allowed the corpus of the explanation dialogues to capture (if there are changes present) how humans structure the explanations based on the role. Subsequent analysis showed that explanation dialogues can be generalised into the same structure and sequence. A grounded approach was employed to code the corpus according to the themes observed. A contrasting approach can be followed to build this explanation corpus, using a controlled study to generate the dialogues [238], though the resulting corpus can have the risk of not generalising

well into other settings.

To be useful for an intelligent agent, an explanation dialogue should have a computational module formalised in an agent understandable manner. To this end, a state model was developed by analysing the frequency and the sequence of the coded concepts. The Agent Dialogue Framework [150] was used to formalise the state model as a dialogue game. Several other frameworks also exist [22, 7, 187] that can be used to formalise the dialogue state model, to describe the interactions between the explainee and the explainer.

An important finding of the RQ4 is the emergence of argumentation within the explanation dialogue. Argumentation was observed at a high frequency across explanation types, and often occurs after an explanation is given. In general, humans can contest an explanation and argue with facts of their own in a natural dialogue. This has implications in developing explanation dialogue systems for XAI systems, as most of the explainable models in the current literature does not provide the ability to contest the explanations. Explanation dialogue models have also been proposed in the social science literature [239, 25, 237], which takes argumentation into account. These models are largely conceptual and suffer from a lack of agent-based formalisation and evaluation in human-agent settings. The explanation dialogue model's coverage was evaluated through a human-agent experiment, having a sequential decision-making agent assisting the participants in the strategy of a board game scenario.

Several works have taken inspiration of a grounded approach in developing explanation dialogues to proposed argumentation frameworks [101], interactive explanations [185, 212, 213] and user-centred explanations [129]. Importantly, these works bridge the gap between the formalised model and the implementation of it in an agent setting. Though the current findings of RQ4 do not directly implement the explanation dialogue model in RL agents, researchers have taken the explanation dialogue concept further to provide sequential explanations in RL agents [251].

7.6 Limitations

The success of an explainable model is tightly coupled to the human factors of the end-users of the system. As accurately modelling the human is a challenging task for current artificial intelligence methods, this naturally also introduces several key limitations that are common to explainable AI systems in general. As the definition of what constitutes an explanation changes

with the context, domain and methods used, distinct limitations are also present in each proposed method [14]. This section discusses the limitations of the explainable models (action influence and distal influence models) proposed in this thesis and limitations in the evaluation methodology that was followed.

7.6.1 Explainable Models

The main drawback of action influence models (and by extension distal explanation models) was the need of a handcrafted causal action influence structure. In Chapter 5, we discussed how this limitation was mitigated through the learning of the influence structure end-to-end just from the RL agent's state-action data. Another limitation lies in handling continuous action spaces, due to the influence structure requiring action labels to annotate the edges of the graph. Recent progress in discretising RL actions [225] can potentially be used as a precursor model to approximately solve this limitation. The level of abstraction explanations are generated can affect the intelligibility of an XAI system. In RL agent explanations, abstraction levels can exist at the action and the state level, and there can be unique action influence models for different levels of abstraction. In Chapter 3, we empirically tested 2 different levels of action influence models, where the less granular model fared better against the other. The number of abstraction levels needed for a domain needs to be determined based on the domain and needs further research.

7.6.2 Evaluation and Environments

RL Environments: Limitations can be introduced through the choice of environments used to evaluate the explainable RL models. In Chapter 3 through 5, we used a combination of OpenAI benchmarks [31] and custom StarCraft II [235] scenarios to computationally asses the performance of the explainable models. There is much debate [123, 114] on the lack of a standard benchmark for RL, introducing inconsistencies in reproducibility challenges in the evaluation. Though we selected different OpenAI environments and StarCraft II scenarios with varying state-action spaces, further evaluations in other domains might be needed.

Human Evaluations: Considering human explanations, the StarCraft II domain was selected for its expressive and visual nature [181], having agents that can exhibit complex strategic

behaviours going beyond path and task planning. Further work is needed in human evaluations, that make use of a different domain in assessing the generalisability of the action influence models.

7.7 Future Work

We propose two main avenues of future directions that stems from the original work presented in this thesis.

7.7.1 Inferring the abstraction level of the explainee

As discussed above, a drawback of the current action influence models is the ambiguity that exists in selecting the correct abstraction level for the explanation. The implementation of multiple explainable models with different abstract levels depends on the granularity of the domain. The selection of the correct abstraction level poses a larger challenge in comparison, in that, the XAI system would need to infer the level needed through the interactions with the explainee. The explanation dialogue model presented in Chapter 6 can be used as the explanation interface to infer the epistemic knowledge and the level of the end-user. Indeed preliminary work has been proposed in this direction by Yeung et al. [251], where interactive sequential explanations for RL agents are used to identify the mental model of the user.

7.7.2 Action influence models beyond explainability

Having a causal model of the environments has the potential to enhance the reasoning capability of intelligent agents [81, 55]. As action influence models function as surrogate models that enable explainability. Surrogate models can also be used for auxiliary tasks. Meta-reasoning and causality based exploration in RL agents are possible future work that can make use of action influence models as surrogate models. Further, as these models are learned at the run-time of the RL agent, exploration of the agents' can be guided based on whether a selected action makes sense based on the causal structure. As an example, the exploration can be guided by 'why' and 'why not' questions and their corresponding explanations, that compare contrasting behaviour.

7.8 Final Remarks

This thesis took inspiration from the large body of literature in cognitive science and philosophy that studies the *nature* of explanation, and introduced causal explainable models for reinforcement learning agents. By following a human-centred approach, our *action influence models* (and subsequent extensions) obtained strong results. The work presented in this thesis paves a way forward in studies of explainability, showing how a fundamental understanding of how humans define explainability can help achieve explainable agency in artificial intelligence.

Bibliography

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. "Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda". In: *Proceedings of the 2018 CHI conference on human factors in computing systems*. ACM. 2018, p. 582.

[2] Suhayya Abu-Hakima and Franz Oppacher. "Improving explanations in knowledge-based systems: RATIONALE". In: *Knowledge Acquisition* 2.4 (1990), pp. 301–343.

[3] Vincent Aleven, Amy Ogan, Octav Popescu, Cristen Torrey, and Kenneth Koedinger. "Evaluating the effectiveness of a tutorial dialogue system for self-explanation". In: *International conference on intelligent tutoring systems*. Springer. 2004, pp. 443–454.

[4] Vincent Aleven, Octav Popescu, and Kenneth R Koedinger. "Towards tutorial dialog to support self-explanation: Adding natural language understanding to a cognitive tutor". In: *Proceedings of Artificial Intelligence in Education*. Citeseer. 2001, pp. 246–255.

[5] Constantin F Aliferis, Ioannis Tsamardinos, and Alexander Statnikov. "HITON: a novel Markov Blanket algorithm for optimal variable selection". In: *AMIA annual symposium proceedings*. Vol. 2003. American Medical Informatics Association. 2003, p. 21.

[6] David Alvarez-Melis and Tommi Jaakkola. "A causal framework for explaining the predictions of black-box sequence-to-sequence models". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 412–421. DOI: 10.18653/v1/D17-1042.

[7] Leila Amgoud and Nabil Hameurlain. "A formal model for designing dialogue strategies". In: *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*. 2006, pp. 414–416.

[8] Leila Amgoud, Nicolas Maudet, and Simon Parsons. "Modelling dialogues using argumentation". In: *Proceedings Fourth International Conference on MultiAgent Systems*. IEEE. 2000, pp. 31–38.

[9] Dan Amir and Ofra Amir. "HIGHLIGHTS: Summarizing Agent Behavior to People". In: *Proc. of the 17th International conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. 2018.

[10] Ofra Amir, Finale Doshi-Velez, and David Sarne. "Summarizing agent strategies". In: *Autonomous Agents and Multi-Agent Systems* 33.5 (2019), pp. 628–644.

[11] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. *Machine Bias*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed: 2020-06-30.

[12] Abdallah Arioua and Madalina Croitoru. "Formalizing explanatory dialogues". In: *International Conference on Scalable Uncertainty Management*. Springer. 2015, pp. 282–297.

[13] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. "A brief survey of deep reinforcement learning". In: *arXiv preprint arXiv:1708.05866* (2017).

[14] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. "One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques". In: *arXiv preprint arXiv:1909.03012* (2019).

[15] Nazia Attari, Martin Heckmann, and David Schlangen. "From explainability to explanation: Using a dialogue setting to elicit annotations with justifications". In: *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*. 2019, pp. 331–335.

[16] Allan Bäck. *Aristotle's theory of abstraction*. Springer, 2014.

[17] Alexander Balke and Judea Pearl. "Counterfactuals and policy analysis in structural models". In: *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc. 1995, pp. 11–18.

[18] Adrian Keith Ball, David C Rye, David Silvera-Tawil, and Mari Velonaki. "How should a robot approach two people?" In: *Journal of Human-Robot Interaction* 6.3 (2017), pp. 71–91.

[19] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, and Antonio Torralba. "Visualizing and Understanding Generative Adversarial Networks". In: *International Conference on Learning Representations*. 2019. URL: https://openreview.net/forum?id=Hyg_X2C5FX.

[20] Trevor JM Bench-Capon, PAUL E Dunne, and Paul H Leng. "Interacting with knowledge-based systems through dialogue games". In: *Proceedings of the Eleventh International Conference on Expert Systems and Applications*. 1991, pp. 123–140.

[21] Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. "A meta-transfer objective for learning to disentangle causal mechanisms". In: *arXiv preprint arXiv:1901.10912* (2019).

[22] Jamal Bentahar, Bernard Moulin, and Brahim Chaib-draa. "Commitment and argument network: a new formalism for agent communication". In: *Workshop on Agent Communication Languages*. Springer. 2003, pp. 146–165.

[23] Michel Besserve, Arash Mehrjou, Rémy Sun, and Bernhard Schölkopf. "Counterfactuals uncover the modular structure of deep generative models". In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=SJxDDpEKvH.

[24] Floris Bex and Douglas Walton. "Combining explanation and argumentation in dialogue". In: *Argument & Computation* 7.1 (2016), pp. 55–68.

[25] Floris Bex and Douglas Walton. "Combining explanation and argumentation in dialogue". In: *Argument & Computation* 7.1 (2016), pp. 55–68.

[26] Alberto Blanco-Justicia, Josep Domingo-Ferrer, Sergio Martinez, and David Sanchez. "Machine learning explainability via microaggregation and shallow decision trees". In: *Knowledge-Based Systems* 194 (2020), p. 105532.

[27] Gisela Böhm and Hans-Rüdiger Pfister. "How people explain their own and others' behavior: a theory of lay causal explanations". In: *Frontiers in psychology* 6 (2015), p. 139.

[28] Kenneth A Bollen. "Structural equation models". In: *Encyclopedia of biostatistics* 7 (2005).

[29] Craig Boutilier, Richard Dearden, Moises Goldszmidt, et al. "Exploiting structure in policy construction". In: *IJCAI*. Vol. 14. 1995, pp. 1104–1113.

[30] Virginia Braun and Victoria Clarke. "Using thematic analysis in psychology". In: *Qualitative research in psychology* 3.2 (2006), pp. 77–101.

[31] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. *OpenAI Gym*. 2016. eprint: arXiv:1606.01540.

[32] Joost Broekens, Maaike Harbers, Koen Hindriks, Karel Van Den Bosch, Catholijn Jonker, and John-Jules Meyer. "Do you get it? User-evaluated explainable BDI agents". In: *German Conference on Multiagent System Technologies*. Springer. 2010, pp. 28–39.

[33] Alexander Brown and Marek Petrik. "Interpretable reinforcement learning with ensemble methods". In: *arXiv preprint arXiv:1809.06995* (2018).

[34] Peter Bühlmann, Markus Kalisch, and Marloes H Maathuis. "Variable selection in high-dimensional linear models: partially faithful distributions and the PC-simple algorithm". In: *Biometrika* 97.2 (2010), pp. 261–278.

[35] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. "Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data?" In: *Perspectives on psychological science* 6.1 (2011), pp. 3–5.

[36] Ruth MJ Byrne. "Counterfactuals in explainable artificial intelligence (XAI): evidence from human reasoning". In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. 2019, pp. 6276–6282.

[37] Alison Cawsey. *Explanation and interaction: the computer generation of explanatory dialogues*. MIT press, 1992.

[38] Alison Cawsey. "Explanatory dialogues". In: *Interacting with computers* 1.1 (1989), pp. 69–92.

[39] Alison Cawsey. "Planning interactive explanations". In: *International Journal of Man-Machine Studies* 38.2 (1993), pp. 169–199.

[40] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. "Plan explanations as model reconciliation: Moving beyond explanation as soliloquy". In: *arXiv preprint arXiv:1701.08317* (2017).

[41] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. "Plan explanations as model reconciliation: Moving beyond explanation as soliloquy". In: *arXiv preprint arXiv:1701.08317* (2017).

[42] Eric Chalmers, Edgar Bermudez Contreras, Brandon Robertson, Artur Luczak, and Aaron Gruber. "Learning to predict consequences as a method of knowledge transfer in reinforcement learning". In: *IEEE transactions on neural networks and learning systems* 29.6 (2017), pp. 2259–2270.

[43] B Chandrasekaran, Michael C Tanner, and John R Josephson. "Explaining control strategies in problem solving". In: *IEEE Intelligent Systems* 1 (1989), pp. 9–15.

[44] Bruce Chandrasekaran, Michael C Tanner, and John R Josephson. "Explaining control strategies in problem solving". In: *IEEE Intelligent Systems* 1 (1989), pp. 9–15.

[45] David Chapman and Leslie Pack Kaelbling. "Input Generalization in Delayed Reinforcement Learning: An Algorithm and Performance Comparisons." In: *IJCAI*. Vol. 91. Citeseer. 1991, pp. 726–731.

[46] Aditya Chattopadhyay, Piyushi Manupriya, Anirban Sarkar, and Vineeth N Balasubramanian. "Neural network attributions: A causal perspective". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 981–990.

[47] Ana Paula Chaves and Marco Aurelio Gerosa. "Single or Multiple Conversational Agents?: An Interactional Coherence Comparison". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM. 2018, p. 191.

[48] David Maxwell Chickering. "Optimal structure identification with greedy search". In: *Journal of machine learning research* 3.Nov (2002), pp. 507–554.

[49] William J Clancey. "The epistemology of a rule-based expert system—a framework for explanation". In: *Artificial intelligence* 20.3 (1983), pp. 215–251.

[50] Diego Colombo and Marloes H Maathuis. "Order-independent constraint-based causal structure learning." In: *J. Mach. Learn. Res.* 15.1 (2014), pp. 3741–3782.

[51] Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. "Learning high-dimensional directed acyclic graphs with latent and selection variables". In: *The Annals of Statistics* (2012), pp. 294–321.

[52] Youri Coppens, Kyriakos Efthymiadis, Tom Lenaerts, and Ann Nowe. "Distilling Deep Reinforcement Learning Policies in Soft Decision Trees". English. In: *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*. Ed. by Tim Miller, Rosina Weber, and Daniele Magazzeni. Aug. 2019, pp. 1–6.

[53] Kenneth James Williams Craik. *The nature of explanation*. Vol. 445. CUP Archive, 1952.

[54] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. "Wizard of Oz studies: why and how". In: *Proceedings of the 1st international conference on Intelligent user interfaces*. 1993, pp. 193–200.

[55] Ishita Dasgupta, Jane Wang, Silvia Chiappa, Jovana Mitrovic, Pedro Ortega, David Raposo, Edward Hughes, Peter Battaglia, Matthew Botvinick, and Zeb Kurth-Nelson. "Causal reasoning from meta-reinforcement learning". In: *arXiv preprint arXiv:1901.08162* (2019).

[56] A Philip Dawid. "Beware of the DAG!" In: *Causality: objectives and assessment*. PMLR. 2010, pp. 59–86.

[57] Maartje M A De Graaf and Bertram F Malle. "How People Explain Action (and Autonomous Intelligent Systems Should Too)". In: *AAAI 2017 Fall Symposium on "AI-HRI"* (2017), pp. 19–26.

[58] Maartje MA De Graaf and Bertram F Malle. "How people explain action (and autonomous intelligent systems should too)". In: *2017 AAAI Fall Symposium Series*. 2017.

[59] Maartje MA De Graaf and Bertram F Malle. "How people explain action (and autonomous intelligent systems should too)". In: *2017 AAAI Fall Symposium Series*. 2017.

[60] Daniel Clement Dennett. *The intentional stance*. MIT press, 1989.

[61] Frank Dignum, Barbara Dunin-Keplicz, and Rineke Verbrugge. "Agent Theory for Team Formation by Dialogue". In: *Intelligent Agents VII Agent Theories Architectures and Languages*. Ed. by Cristiano Castelfranchi and Yves Lespérance. Springer Berlin Heidelberg, 2001, pp. 150–166.

[62] Zihan Ding, Pablo Hernandez-Leal, Gavin Weiguang Ding, Changjian Li, and Ruitong Huang. "CDT: Cascading Decision Trees for Explainable Reinforcement Learning". In: *arXiv preprint arXiv:2011.07553* (2020).

[63] Finale Doshi-Velez and Been Kim. "Towards a rigorous science of interpretable machine learning". In: *arXiv preprint arXiv:1702.08608* (2017).

[64] Phil Dowe. "Wesley Salmon's process theory of causality and the conserved quantity theory". In: *Philosophy of science* 59.2 (1992), pp. 195–216.

[65] Gengshen Du, Michael M Richter, and Guenther Ruhe. "An Explanation Oriented Dialogue Approach for Solving Wicked Planning Problems." In: *ExaCt*. 2005, pp. 62–75.

[66] Mateusz Dubiel. "Towards Human-Like Conversational Search Systems". In: *Proceedings of the 2018 Conference on Human Information Interaction&Retrieval*. ACM. 2018, pp. 348–350.

[67] Francisco Elizalde, L Enrique Sucar, Manuel Luque, J Diez, and Alberto Reyes. "Policy explanation in factored Markov decision processes". In: *Proceedings of the 4th European Workshop on Probabilistic Graphical Models (PGM 2008)*. 2008, pp. 97–104.

[68] Francisco Elizalde and Luis Enrique Sucar. "Expert Evaluation of Probabilistic Explanations." In: *ExaCt*. 2009, pp. 1–12.

[69] Francisco Elizalde, Luis Enrique Sucar, Alberto Reyes, and Pablo Debuen. "An MDP Approach for Explanation Generation." In: *ExaCt*. 2007, pp. 28–33.

[70] Tom Everitt, Ramana Kumar, Victoria Krakovna, and Shane Legg. "Modeling AGI safety frameworks with causal influence diagrams". In: *arXiv preprint arXiv:1906.08663* (2019).

[71] Tom Everitt, Pedro A Ortega, Elizabeth Barnes, and Shane Legg. "Understanding Agent Incentives using Causal Influence Diagrams. Part I: Single Action Settings". In: *arXiv preprint arXiv:1902.09980* (2019).

[72] *Explainable AI | Google Cloud*. https://cloud.google.com/explainable-ai. Accessed: 2020-06-30.

[73] *Explainable AI Is A Game-Changer For Business Analytics*. https://www.forbes.com/sites/forbestechcouncil/2020/06/08/explainable-ai-is-a-game-changer-for-business-analytics/#5268ba2a5ea1. Accessed: 2020-06-30.

[74] Andrea Falcon. "Aristotle on causality". In: (2006).

[75] Maria Fox, Derek Long, and Daniele Magazzeni. "Explainable Planning". In: *IJCAI - Workshop on Explainable AI* (2017).

[76] Maria Fox, Derek Long, and Daniele Magazzeni. "Explainable planning". In: *arXiv preprint arXiv:1709.10256* (2017).

[77] Yosuke Fukuchi, Masahiko Osawa, Hiroshi Yamakawa, and Michita Imai. "Autonomous self-explanation of behavior for interactive reinforcement learning agents". In: *Proceedings of the 5th International Conference on Human Agent Interaction*. ACM. 2017, pp. 97–101.

[78] Ya'akov Gal and Avi Pfeffer. "Networks of influence diagrams: A formalism for representing agents' beliefs and decision-making processes". In: *Journal of Artificial Intelligence Research* 33 (2008), pp. 109–147.

[79] Dan Geiger and David Heckerman. "Learning gaussian networks". In: *Uncertainty Proceedings 1994*. Elsevier, 1994, pp. 235–243.

[80] Dan Geiger, Thomas Verma, and Judea Pearl. "d-separation: From theorems to algorithms". In: *Machine Intelligence and Pattern Recognition*. Vol. 10. Elsevier, 1990, pp. 139–148.

[81] Samuel J Gershman. "Reinforcement learning and causal models". In: *The Oxford handbook of causal reasoning* (2017), p. 295.

[82] Nigel Gilbert. "Explanation and dialogue". In: *The Knowledge Engineering Review* 4.3 (1989), pp. 235–247.

[83] Barney G Glaser, Anselm L Strauss, and Elizabeth Strutzel. "The discovery of grounded theory; strategies for qualitative research". In: *Nursing research* 17.4 (1968), p. 364.

[84] *Google tackles the black box problem with Explainable AI*. https://www.bbc.com/news/technology-50506431. Accessed: 2020-06-30.

[85] Alison Gopnik, Clark Glymour, David M Sobel, Laura E Schulz, Tamar Kushnir, and David Danks. "A theory of causal learning in children: causal maps and Bayes nets." In: *Psychological review* 111.1 (2004), p. 3.

[86] Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, Isabelle Guyon, David Lopez-Paz, and Michele Sebag. "Learning functional causal models with generative neural networks". In: *Explainable and interpretable models in computer vision and machine learning*. Springer, 2018, pp. 39–80.

[87] Clive WJ Granger. "Investigating causal relations by econometric models and cross-spectral methods". In: *Econometrica: journal of the Econometric Society* (1969), pp. 424–438.

[88] Samuel Greydanus, Anurag Koul, Jonathan Dodge, and Alan Fern. "Visualizing and understanding atari agents". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 1792–1801.

[89] David Gunning and David W Aha. "DARPA's Explainable Artificial Intelligence Program". In: *AI Magazine* 40.2 (2019), pp. 44–58.

[90] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. "XAI—Explainable artificial intelligence". In: *Science Robotics* 4.37 (2019).

[91] Joseph Y Halpern and Judea Pearl. "Causes and explanations: A structural-model approach. Part I: Causes". In: *The British journal for the philosophy of science* 56.4 (2005), pp. 843–887.

[92] Joseph Y Halpern and Judea Pearl. "Causes and explanations: A structural-model approach. Part II: Explanations". In: *The British journal for the philosophy of science* 56.4 (2005), pp. 889–911.

[93] Joseph Y. Halpern and Judea Pearl. "Causes and Explanations: A Structural-Model Approach-Part II: Explanations". In: *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 1*. IJCAI'01. Seattle, WA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 27–34. ISBN: 1558608125.

[94] Robert James Hankinson. *Cause and explanation in ancient Greek thought*. Oxford University Press, 2001.

[95] Michael Harradon, Jeff Druce, and Brian Ruttenberg. "Causal learning and explanation of deep neural networks via autoencoded activations". In: *arXiv preprint arXiv:1802.00541* (2018).

[96] Diane Warner Hasling, William J Clancey, and Glenn Rennels. "Strategic explanations for a diagnostic consultation system". In: *International Journal of Man-Machine Studies* 20.1 (1984), pp. 3–19.

[97] Alain Hauser and Peter Bühlmann. "Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs". In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 2409–2464.

[98] Leslie Hayduk, Greta Cummings, Rainer Stratkotter, Melanie Nimmo, Kostyantyn Grygoryev, Donna Dosman, Michael Gillespie, Hannah Pazderka-Robinson, and Kwame Boadu. "Pearl's D-separation: One more step into causal thinking". In: *Structural Equation Modeling* 10.2 (2003), pp. 289–311.

[99] Bradley Hayes and Julie A Shah. "Improving robot controller transparency through autonomous policy explanation". In: *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*. ACM. 2017, pp. 303–312.

[100] David Heckerman, Dan Geiger, and David M Chickering. "Learning Bayesian networks: The combination of knowledge and statistical data". In: *Machine learning* 20.3 (1995), pp. 197–243.

[101] Clément Henin and Daniel Le Métayer. "Towards a framework for challenging ML-based decisions". In: *1st International Workshop on Deceptive AI@ ECAI2020*. 2020.

[102] Germund Hesslow. "The problem of causal selection". In: *Contemporary science and natural explanation: Commonsense conceptions of causality* (1988), pp. 11–32.

[103] Alexandre Heuillet, Fabien Couthouis, and Natalia Dıaz-Rodrıguez. "Explainability in deep reinforcement learning". In: *Knowledge-Based Systems* 214 (2021), p. 106685.

[104] Denis Hilton. "Causal explanation". In: *Social psychology: Handbook of basic principles* (2007), pp. 232–253.

[105] Denis J Hilton. "A conversational model of causal explanation". In: *European review of social psychology* 2.1 (1991), pp. 51–81.

[106] Denis J Hilton. "Conversational processes and causal explanation." In: *Psychological Bulletin* 107.1 (1990), p. 65.

[107] Denis J Hilton and L McCLURE John. "The course of events: counterfactuals, causal sequences, and explanation". In: *The Psychology of Counterfactual Thinking*. Routledge, 2007, pp. 56–72.

[108] Denis J Hilton, John McClure, and Robbie M Sutton. "Selecting explanations from causal chains: Do statistical principles explain preferences for voluntary causes?" In: *European Journal of Social Psychology* 40.3 (2010), pp. 383–400.

[109] Denis J Hilton, John L McClure, and Ben R. Slugoski. "The course of events: counterfactuals, causal sequences, and explanation". In: *The Psychology of Counterfactual Thinking*. Ed. by D.R. Mandel, D.J. Hilton, and P. Catellani. Routledge, 2005, pp. 56–72.

[110] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. "Metrics for Explainable AI: Challenges and Prospects". In: *arXiv preprint arXiv:1812.04608* (2018).

[111] Jennifer Hornsby. "Agency and causal explanation." In: (1993).

[112] Jianfeng Huang, Plamen P Angelov, and Chengliang Yin. "Interpretable policies for reinforcement learning by empirical fuzzy sets". In: *Engineering Applications of Artificial Intelligence* 91 (2020), p. 103559.

[113] Xiaoshui Huang, Fujin Zhu, Lois Holloway, and Ali Haidar. "Causal Discovery from Incomplete Data using An Encoder and Reinforcement Learning". In: *arXiv preprint arXiv:2006.05554* (2020).

[114] Scott Jordan, Yash Chandak, Daniel Cohen, Mengxue Zhang, and Philip Thomas. "Evaluating the performance of reinforcement learning algorithms". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 4962–4973.

[115] John R Josephson and Susan G Josephson. *Abductive inference: Computation, philosophy, technology*. Cambridge University Press, 1996.

[116] Zoe Juozapaitis, Anurag Koul, Alan Ferm, Martin Erwig, and Finale Doshi-Velez. "Explainable Reinforcement Learning via Reward Decomposition". English. In: *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*. Ed. by Tim Miller, Rosina Weber, and Daniele Magazzeni. Aug. 2019, pp. 47–53.

[117] Zoe Juozapaitis, Anurag Koul, Alan Fern, Martin Erwig, and Finale Doshi-Velez. "Explainable reinforcement learning via reward decomposition". In: *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*. 2019, pp. 47–53.

[118] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. "Reinforcement learning: A survey". In: *Journal of artificial intelligence research* 4 (1996), pp. 237–285.

[119] Diviyan Kalainathan, Olivier Goudet, Isabelle Guyon, David Lopez-Paz, and Michèle Sebag. "Sam: Structural agnostic model, causal discovery and penalized adversarial learning". In: (2018).

[120] Robert Kass and Tim Finin. *The Need for User Models in Generating Expert System Explanations*. Tech. rep. University of Pennsylvania, June 1988, pp. 1–32.

[121] Gideon Keren. "Between-or within-subjects design: A methodological dilemma". In: *A Handbook for Data Analysis in the Behaviorial Sciences* 1 (2014), pp. 257–272.

[122] Omar Zia Khan, Pascal Poupart, and James P Black. "Minimal Sufficient Explanations for Factored Markov Decision Processes." In: *ICAPS*. 2009.

[123] Khimya Khetarpal, Zafarali Ahmed, Andre Cianflone, Riashat Islam, and Joelle Pineau. "Re-evaluate: Reproducibility in evaluating reinforcement learning algorithms". In: (2018).

[124] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015. URL: http://arxiv.org/abs/1412.6980.

[125] Gary Klein. "Explaining explanation, part 3: The causal landscape". In: *IEEE Intelligent Systems* 33.2 (2018), pp. 83–88.

[126] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012), pp. 1097–1105.

[127] Erich Kummerfeld, David Danks, and M Cognition. "Online Learning of Time-varying Causal Structures". In: (2012).

[128] Joan Kung. "Aristotle on essence and explanation". In: *Philosophical Studies* 31.6 (1977), pp. 361–383.

[129] Michał Kuźba and Przemysław Biecek. "What Would You Ask the Machine Learning Model? Identification of User Needs for Model Explanations Based on Human-Model Conversations". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2020, pp. 447–459.

[130] Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. "Gradient-Based Neural DAG Learning". In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=rklbKA4YDS.

[131] Pat Langley, Ben Meadows, Mohan Sridharan, and Dongkyu Choi. "Explainable agency for intelligent autonomous systems". In: *Twenty-Ninth IAAI Conference*. 2017.

[132] Thuc Duy Le, Tao Hoang, Jiuyong Li, Lin Liu, Huawen Liu, and Shu Hu. "A fast PC algorithm for high dimensional causal discovery with multi-core PCs". In: *IEEE/ACM transactions on computational biology and bioinformatics* 16.5 (2016), pp. 1483–1495.

[133] David Lewis. "Causation". In: *The journal of philosophy* 70.17 (1974), pp. 556–567.

[134] David K. Lewis. "Causal Explanation". In: *Philosophical Papers Vol. Ii*. Ed. by David Lewis. Oxford University Press, 1986, pp. 214–240.

[135] Brian Y Lim, Anind K Dey, and Daniel Avrahami. "Why and why not explanations improve the intelligibility of context-aware intelligent systems". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2009, pp. 2119–2128.

[136] Long-Ji Lin. "Self-improving reactive agents based on reinforcement learning, planning and teaching". In: *Machine learning* 8.3-4 (1992), pp. 293–321.

[137] Michael P Linegang, Heather A Stoner, Michael J Patterson, Bobbie D Seppelt, Joshua D Hoffman, Zachariah B Crittendon, and John D Lee. "Human-automation collaboration in dynamic mission planning: A challenge requiring an ecological approach". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 50. 23. SAGE Publications Sage CA: Los Angeles, CA. 2006, pp. 2482–2486.

[138] Peter Lipton. "Contrastive explanation". In: *Royal Institute of Philosophy Supplements* 27 (1990), pp. 247–266.

[139] Tania Lombrozo. "Explanation and abductive inference." In: (2012).

[140] Tania Lombrozo. "Simplicity and probability in causal explanation". In: *Cognitive psychology* 55.3 (2007), pp. 232–257.

[141] Tania Lombrozo. "The structure and function of explanations". In: *Trends in cognitive sciences* 10.10 (2006), pp. 464–470.

[142] Tania Lombrozo and Nadya Vasilyeva. "Causal explanation". In: *Oxford handbook of causal reasoning* (2017), pp. 415–432.

[143] Crisrael Lucero, Christianne Izumigawa, Kurt Frederiksen, Lena Nans, Rebecca Iden, and Douglas S Lange. "Human-Autonomy Teaming and Explainable AI Capabilities in RTS Games". In: *International Conference on Human-Computer Interaction*. Springer. 2020, pp. 161–171.

[144] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. "From local explanations to global understanding with explainable AI for trees". In: *Nature machine intelligence* 2.1 (2020), pp. 56–67.

[145] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. "A Grounded Interaction Protocol for Explainable Artificial Intelligence". In: *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. 2019, pp. 1033–1041.

[146] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. "Explainable Reinforcement Learning Through a Causal Lens". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2020.

[147] Karthik Mahadevan, Sowmya Somanath, and Ehud Sharlin. "Communicating Awareness and Intent in Autonomous Vehicle-Pedestrian Interaction". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM. 2018, p. 429.

[148] Ričards Marcinkevičs and Julia E Vogt. "Interpretable Models for Granger Causality Using Self-explaining Neural Networks". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=DEa4JdMWRHp.

[149] Alvaro Parafita Martinez and Jordi Vitria Marca. "Explaining Visual Models by Causal Attribution". In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE. 2019, pp. 4167–4175.

[150] Peter McBurney and Simon Parsons. "Games that agents play: A formal framework for dialogues between autonomous agents". In: *Journal of logic, language and information* 11.3 (2002), pp. 315–334.

[151] John McClure and Denis Hilton. "For you can't always get what you want: When preconditions are better explanations than goals". In: *British Journal of Social Psychology* 36.2 (1997), pp. 223–240.

[152] John McClure, Denis J Hilton, and Robbie M Sutton. "Judgments of voluntary and physical causes in causal chains: Probabilistic and social functionalist criteria for attributions". In: *European journal of social psychology* 37.5 (2007), pp. 879–901.

[153] Christopher Meek. "Graphical Models: Selecting causal and statistical models". PhD thesis. PhD thesis, Carnegie Mellon University, 1997.

[154] Luke Metz, Julian Ibarz, Navdeep Jaitly, and James Davidson. "Discrete sequential prediction of continuous actions for deep rl". In: *arXiv preprint arXiv:1705.05035* (2017).

[155] Tim Miller. "Contrastive Explanation: A Structural-Model Approach". In: *arXiv preprint arXiv:1811.03163* (2018).

[156] Tim Miller. "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial Intelligence* (2018).

[157] Tim Miller. "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial Intelligence* 267 (2019), pp. 1–38.

[158] Tim Miller, Piers Howe, and Liz Sonenberg. "Explainable AI: Beware of Inmates Running the Asylum; Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences". In: *IJCAI-17 Workshop on Explainable AI (XAI)*. 2017, p. 36.

[159] Vibhu O Mittal and Cécile L Paris. "Generating explanations in context: The system perspective". In: *Expert Systems with Applications* 8.4 (1995), pp. 491–503.

[160] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. "Human-level control through deep reinforcement learning". In: *Nature* 518.7540 (2015), p. 529.

[161] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[162] Johanna D Moore and Cécile L Paris. "Requirements for an expert system explanation facility". In: *Computational Intelligence* 7.4 (1991), pp. 367–370.

[163] Johanna D Moore and William R Swartout. "Pointing: A Way Toward Explanation Dialogue." In: *AAAI*. Vol. 90. 1990, pp. 457–464.

[164] Raha Moraffah, Mansooreh Karami, Ruocheng Guo, Adrienne Raglin, and Huan Liu. "Causal interpretability for machine learning-problems, methods and evaluation". In: *ACM SIGKDD Explorations Newsletter* 22.1 (2020), pp. 18–33.

[165] Jonas Nagel and Simon Stephan. "Explanations in causal chains: Selecting distal causes requires exportable mechanisms". In: *Proceedings of the 38th Annual Conference of the Cognitive Science Society*. Cognitive Science Society Austin, TX. 2016, pp. 806–812.

[166] Tanmayee Narendra, Anush Sankaran, Deepak Vijaykeerthy, and Senthil Mani. "Explaining deep learning models using causal inference". In: *arXiv preprint arXiv:1811.04376* (2018).

[167] Meike Nauta, Doina Bucur, and Christin Seifert. "Causal discovery with attention-based convolutional neural networks". In: *Machine Learning and Knowledge Extraction* 1.1 (2019), pp. 312–340.

[168] Joshua Newn, Fraser Allison, Eduardo Velloso, and Frank Vetere. "Looks can be deceiving: Using gaze visualisation to predict and mislead opponents in strategic gameplay". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 2018, pp. 1–12.

[169] Ignavier Ng, Shengyu Zhu, Zhitang Chen, and Zhuangyan Fang. "A Graph Autoencoder Approach to Causal Structure Learning". In: *CoRR* abs/1911.07420 (2019). URL: http://arxiv.org/abs/1911.07420.

[170] Ulf Nielsen, Jean-Philippe Pellet, and André Elisseeff. "Explanation trees for causal Bayesian networks". In: *arXiv preprint arXiv:1206.3276* (2012).

[171] Nils J Nilsson et al. "Shakey the robot". In: (1984).

[172] F. Nothdurft, G. Bertrand, Helmut Lang, and W. Minker. "Adaptive Explanation Architecture for Maintaining Human-Computer Trust". In: *2012 IEEE 36th Annual Computer Software and Applications Conference* (2012), pp. 176–184.

[173] Florian Nothdurft, Tobias Heinroth, and Wolfgang Minker. "The impact of explanation dialogues on human-computer trust". In: *International Conference on Human-Computer Interaction*. Springer. 2013, pp. 59–67.

[174] Ana Rodriguez Palmeiro, Sander van der Kint, Luuk Vissers, Haneen Farah, Joost CF de Winter, and Marjan Hagenzieker. "Interaction between pedestrians and automated vehicles: A Wizard of Oz experiment". In: *Transportation research part F: traffic psychology and behaviour* 58 (2018), pp. 1005–1020.

[175] Judea Pearl. "Causal diagrams for empirical research". In: *Biometrika* 82.4 (1995), pp. 669–688.

[176] Judea Pearl et al. "Causal inference in statistics: An overview". In: *Statistics surveys* 3 (2009), pp. 96–146.

[177] Judea Pearl. *Causality*. Cambridge university press, 2009.

[178] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

[179] Sean Penney, Jonathan Dodge, Claudia Hilderbrand, Andrew Anderson, Logan Simpson, and Margaret Burnett. "Toward Foraging for Understanding of StarCraft Agents: An Empirical Study". In: *23rd International Conference on Intelligent User Interfaces*. ACM. 2018, pp. 225–237.

[180] Sundar Pichai. *AI at Google: our principles*. https://www.blog.google/technology/ai/ai-principles/. Accessed: 2020-06-30.

[181] Rey Pocius, Lawrence Neal, and Alan Fern. "Strategic tasks for explainable reinforcement learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 10007–10008.

[182] Athanasios S Polydoros and Lazaros Nalpantidis. "Survey of model-based reinforcement learning: Applications on robotics". In: *Journal of Intelligent & Robotic Systems* 86.2 (2017), pp. 153–173.

[183] Henry Prakken. "Formal systems for persuasion dialogue". In: *Knowledge Engineering Review* 21.2 (2006), p. 163.

[184] Henry Prakken and Giovanni Sartor. "Modelling Reasoning with Precedents in a Formal Dialogue Game". In: *Judicial Applications of Artificial Intelligence*. Ed. by Giovanni Sartor and Karl Branting. Springer Netherlands, 1998, pp. 127–183.

[185] Antonio Rago, Oana Cocarascu, Christos Bechlivanidis, and Francesca Toni. "Argumentation as a framework for interactive explanations for recommendations". In: *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*. Vol. 17. 1. 2020, pp. 805–815.

[186] Joseph Ramsey, Madelyn Glymour, Ruben Sanchez-Romero, and Clark Glymour. "A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images". In: *International journal of data science and analytics* 3.2 (2017), pp. 121–129.

[187] Chris Reed. "Dialogue frames in agent communication". In: *Proceedings International Conference on Multi Agent Systems (Cat. No. 98EX160)*. IEEE. 1998, pp. 246–253.

[188] *Responsible AI Practices*. https://ai.google/responsibilities/responsible-ai-practices/?category=interpretability. Accessed: 2020-06-30.

[189] Alexander Reutlinger and Juha Saatsi. *Explanation beyond causation: Philosophical perspectives on non-causal explanations*. Oxford University Press, 2018.

[190] Alexander Reutlinger and Juha Saatsi. *Explanation beyond causation: Philosophical perspectives on non-causal explanations*. Oxford University Press, 2018.

[191] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[192] Thomas Richardson, Peter Spirtes, et al. "Ancestral graph Markov models". In: *The Annals of Statistics* 30.4 (2002), pp. 962–1030.

[193] James A Rosenblum and Johanna D Moore. "Participating in instructional dialogues: Finding and exploiting relevant prior explanations". In: *Proceedings of the World Conference on Artificial Intelligence in Education*. Citeseer. 1993, pp. 145–152.

[194] Aaron M Roth, Nicholay Topin, Pooyan Jamshidi, and Manuela Veloso. "Conservative Q-Improvement: Reinforcement Learning for an Interpretable Decision-Tree Policy". In: *arXiv preprint arXiv:1907.01180* (2019).

[195] Donald B Rubin. "Estimating causal effects of treatments in randomized and nonrandomized studies." In: *Journal of educational Psychology* 66.5 (1974), p. 688.

[196] Cynthia Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature Machine Intelligence* 1.5 (2019), pp. 206–215.

[197] Wesley C Salmon. *Scientific explanation and the causal structure of the world*. Princeton University Press, 1984.

[198] I. Sassoon, Nadin Kökciyan, E. Sklar, and Simon Parsons. "Explainable Argumentation for Wellness Consultation". In: *EXTRAAMAS@AAMAS*. 2019.

[199] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. "The graph neural network model". In: *IEEE transactions on neural networks* 20.1 (2008), pp. 61–80.

[200] RP Schank. *Explanation patterns: Understanding mechanically and creatively*. Psychology Press, 2013.

[201] Richard Scheines. "An introduction to causal inference". In: (1997).

[202] Mike Schuster and Kuldip K Paliwal. "Bidirectional recurrent neural networks". In: *IEEE Transactions on Signal Processing* 45.11 (1997), pp. 2673–2681.

[203] Patrick Schwab and Walter Karlen. "CXPlain: Causal explanations for model interpretation under uncertainty". In: *Advances in Neural Information Processing Systems*. 2019, pp. 10220–10230.

[204] Gideon Schwarz et al. "Estimating the dimension of a model". In: *Annals of statistics* 6.2 (1978), pp. 461–464.

[205] Pedro Sequeira and Melinda Gervasio. "Interestingness elements for explainable reinforcement learning: Understanding agents' capabilities and limitations". In: *Artificial Intelligence* 288 (2020), p. 103367.

[206] Eyal Shulman and Lior Wolf. "Meta Decision Trees for Explainable Recommendation Systems". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020, pp. 365–371.

[207] Ronal Singh, Tim Miller, Joshua Newn, Liz Sonenberg, Eduardo Velloso, and Frank Vetere. "Combining planning with gaze for online human intention recognition". In: *Proceedings of the 17th international conference on autonomous agents and multiagent systems*. 2018, pp. 488–496.

[208] Elizabeth I Sklar and Mohammad Q Azhar. "Explanation through argumentation". In: *Proceedings of the 6th International Conference on Human-Agent Interaction*. 2018, pp. 277–285.

[209] Bradford Skow. "Are there non-causal explanations (of particular events)?" In: *The British Journal for the Philosophy of Science* (2020).

[210] Steven Sloman. *Causal models: How people think about the world and its alternatives*. Oxford University Press, 2005.

[211] Susan A Slotnick and Johanna D Moore. "Explaining quantitative systems to uninitiated users". In: *Expert Systems with Applications* 8.4 (1995), pp. 475–490.

[212] Kacper Sokol and Peter Flach. "Limetree: Interactively customisable explanations based on local surrogate multi-output regression trees". In: *arXiv preprint arXiv:2005.01427* (2020).

[213] Kacper Sokol and Peter Flach. "One explanation does not fit all". In: *KI-Künstliche Intelligenz* 34.2 (2020), pp. 235–250.

[214] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

[215] Oliver Struckmeier, Mattia Racca, and Ville Kyrki. "Autonomous Generation of Robust and Focused Explanations for Robot Policies". In: *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2019, pp. 1–8.

[216] Gail M Sullivan and Anthony R Artino Jr. "Analyzing and interpreting data from Likert-type scales". In: *Journal of graduate medical education* 5.4 (2013), pp. 541–542.

[217] Dicky Suryadi and Piotr J Gmytrasiewicz. "Learning models of other agents using influence diagrams". In: *UM99 User Modeling*. Springer, 1999, pp. 223–232.

[218] Daniel Suthers, Beverly Woolf, and Matthew Cornell. "Steps from explanation planning to model construction dialogues". In: *AAAI*. 1992, pp. 24–30.

[219] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[220] Richard S Sutton, David A McAllester, Satinder P Singh, Yishay Mansour, et al. "Policy gradient methods for reinforcement learning with function approximation." In: *NIPs*. Vol. 99. Citeseer. 1999, pp. 1057–1063.

[221] W. Swartout. "Explaining and Justifying Expert Consulting Programs". In: *IJCAI*. 1981.

[222] William R Swartout. "A digitalis therapy advisor with explanations". In: *Proceedings of the 5th international joint conference on Artificial intelligence-Volume 2*. 1977, pp. 819–825.

[223] William R Swartout. "XPLAIN: A system for creating and explaining expert consulting programs". In: *Artificial intelligence* 21.3 (1983), pp. 285–325.

[224] Aaquib Tabrez, Shivendra Agrawal, and Bradley Hayes. "Explanation-based reward coaching to improve human performance via reinforcement learning". In: *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2019, pp. 249–257.

[225] Yunhao Tang and Shipra Agrawal. "Discretizing continuous action space for on-policy optimization". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 04. 2020, pp. 5981–5988.

[226] Julia Tanney. "Reasons as non-causal, context-placing explanations". In: *New essays on the explanation of action*. Springer, 2009, pp. 94–111.

[227] Nicholay Topin and Manuela Veloso. "Generation of policy-level explanations for reinforcement learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 2019, pp. 2514–2521.

[228] Michail Tsagris. "Bayesian network learning with the PC algorithm: An improved and correct variation". In: *Applied Artificial Intelligence* 33.2 (2019), pp. 101–123.

[229] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. "The max-min hill-climbing Bayesian network structure learning algorithm". In: *Machine learning* 65.1 (2006), pp. 31–78.

[230] Jeroen Van Bouwel and Erik Weber. "Remote causes, bad explanations?" In: *Journal for the Theory of Social Behaviour* 32.4 (2002), pp. 437–449.

[231] Marko Vasic, Andrija Petrovic, Kaiyuan Wang, Mladen Nikolic, Rishabh Singh, and Sarfraz Khurshid. "Moët: Interpretable and verifiable reinforcement learning via mixture of expert trees". In: *arXiv preprint arXiv:1906.06717* (2019).

[232] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[233] Abhinav Verma, Vijayaraghavan Murali, Rishabh Singh, Pushmeet Kohli, and Swarat Chaudhuri. "Programmatically interpretable reinforcement learning". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 5045–5054.

[234] Serena Villata, Guido Boella, Dov M Gabbay, and Leendert Van Der Torre. "A socio-cognitive model of trust using argumentation theory". In: *International Journal of Approximate Reasoning* 54.4 (2013), pp. 541–559.

[235] Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani, Heinrich Küttler, John Agapiou, Julian Schrittwieser, et al. "Starcraft ii: A new challenge for reinforcement learning". In: *arXiv preprint arXiv:1708.04782* (2017).

[236] Jasper van der Waa, Jurriaan van Diggelen, Karel van den Bosch, and Mark Neerincx. "Contrastive explanations for reinforcement learning in terms of expected consequences". In: *arXiv preprint arXiv:1807.08706* (2018).

[237] Douglas Walton. "A dialogue system for evaluating explanations". In: *Argument evaluation and evidence*. Springer, 2016, pp. 69–116.

[238] Douglas Walton. "A dialogue system specification for explanation". In: *Synthese* 182.3 (2011), pp. 349–374.

[239] Douglas Walton. "Dialogical Models of Explanation." In: *ExaCt* 2007 (2007), pp. 1–9.

[240] Douglas Walton and Erik C. W. Krabbe. *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. State University of New York Press, 1995.

[241] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. "Designing Theory-Driven User-Centric Explainable AI". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM. 2019, p. 601.

[242] Jianhong Wang, Yuan Zhang, Tae-Kyun Kim, and Yunjie Gu. "Shapley Q-value: a local reward approach to solve global reward games". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05. 2020, pp. 7285–7292.

[243] Ning Wang, David V Pynadath, and Susan G Hill. "Trust calibration within a human-robot team: Comparing automatically generated explanations". In: *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press. 2016, pp. 109–116.

[244] Weixun Wang, Tianpei Yang, Yong Liu, Jianye Hao, Xiaotian Hao, Yujing Hu, Yingfeng Chen, Changjie Fan, and Yang Gao. "Action Semantics Network: Considering the Effects of Actions in Multiagent Systems". In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=ryg48p4tPH.

[245] Michael R. Wick and James R. Slagle. "An explanation facility for today's expert systems". In: *IEEE Expert* 4.1 (1989), pp. 26–36.

[246] Michael Winikoff. "Debugging Agent Programs with Why?: Questions". In: *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. AAMAS '17. IFAAMAS, 2017, pp. 251–259.

[247] James Woodward. "Counterfactuals and causal explanation". In: *International Studies in the Philosophy of Science* 18.1 (2004), pp. 41–72.

[248] James Woodward. *Making things happen: A theory of causal explanation*. Oxford university press, 2005.

[249] Guandong Xu, Tri Dung Duong, Qian Li, Shaowu Liu, and Xianzhi Wang. "Causality Learning: A New Perspective for Interpretable Machine Learning". In: *arXiv preprint arXiv:2006.16789* (2020).

[250] L Richard Ye and Paul E Johnson. "The impact of explanation facilities on user acceptance of expert systems advice". In: *Mis Quarterly* (1995), pp. 157–172.

[251] Arnold Yeung, Shalmali Joshi, Joseph Jay Williams, and Frank Rudzicz. "Sequential Explanations with Mental Model-Based Policies". In: *arXiv preprint arXiv:2007.09028* (2020).

[252] Yue Yu, Jie Chen, Tian Gao, and Mo Yu. "DAG-GNN: DAG Structure Learning with Graph Neural Networks". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, Sept. 2019, pp. 7154–7163. URL: http://proceedings.mlr.press/v97/yu19a.html.

[253] Zhiwei Zeng, C. Miao, C. Leung, and J. Chin. "Building More Explainable Artificial Intelligence With Argumentation". In: *AAAI*. 2018.

[254] Junzhe Zhang and Elias Bareinboim. "Fairness in Decision-Making–The Causal Explanation Formula". In: *32nd AAAI Conference on Artificial Intelligence*. 2018.

[255] Qingyuan Zhao and Trevor Hastie. "Causal interpretations of black-box models". In: *Journal of Business & Economic Statistics* 39.1 (2021), pp. 272–281.

[256] Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. "DAGs with NO TEARS: Continuous Optimization for Structure Learning". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS'18. Montréal, Canada: Curran Associates Inc., 2018, pp. 9492–9503.

[257] Shengyu Zhu, Ignavier Ng, and Zhitang Chen. "Causal Discovery with Reinforcement Learning". In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=S1g2skStPB.